FUNDAMENTOS MATEMÁTICOS DE REGRESIÓN LINEAL

Parte II

Moisés Arreguín Sámano

Andrea Damaris Hernández Allauca

Benedicto Vargas Larreta

Juan Enrique Ureña Moreno

CASA EDITORA DEL POLO

FUNDAMENTOS MATEMÁTICOS DE REGRESIÓN LINEAL

Parte II

Moisés Arreguín- Sámano, Docente de la Universidad Estatal de Bolívar, Bolívar, Ecuador.

Andrea Damaris Hernández- Allauca, Docente Investigador Escuela Superior Politécnica de Chimborazo- Grupo de Investigación y Transferencia de Tecnologías en Recursos Hídricos (GITRH) y Grupo de Investigación en Turismo (GITUR). Riobamba, Ecuador.

Benedicto Vargas- Larreta, Doctor en Ciencias por la Universidad de Gotinga, Alemania. Profesor-investigador en el Tecnológico Nacional de México Campus El Salto, Durango, donde imparte cursos como manejo forestal, evaluación de recursos forestales, silvicultura y biometría forestal.

Juan Enrique Ureña- Moreno, Ingeniero Civil por la Pontificia Universidad Católica del Ecuador, Master en Ciudades y Arquitectura Sostenible por la Universidad de Sevilla, España, Docente en la Escuela Superior Politécnica de Chimborazo.

Casa Editora del Polo - CASEDELPO CIA. LTDA.
Departamento de Edición

Editado y distribuido por:

Editorial: Casa Editora del Polo
Sello Editorial: 978-9942-816
Manta, Manabí, Ecuador. 2019
Teléfono: (05) 6051775 / 0991871420
Web: www.casedelpo.com
ISBN: 978-9942-621-01-6

Revisión, Ortografía y Redacción:
Lic. Jessica Mero Vélez

Diseño de Portada:
Michael Josué Suárez-Espinar

Diagramación:
Ing. Edwin Alejandro Delgado-Veliz

Director Editorial:
Dra. Tibisay Milene Lamus-García

Todos los libros publicados por la Casa Editora del Polo, son sometidos previamente a un proceso de evaluación realizado por árbitros calificados.

Este es un libro digital y físico, destinado únicamente al uso personal y colectivo en trabajos académicos de investigación, docencia y difusión del Conocimiento, donde se debe brindar crédito de manera adecuada a los autores.

Comité Científico Académico

Constancia de Arbitraje

La Casa Editora del Polo, hace constar que este libro proviene de una investigación realizada por los autores, siendo sometido a un arbitraje bajo el sistema de doble ciego (peer review), de contenido y forma por jurados especialistas. Además, se realizó una revisión del enfoque, paradigma y método investigativo; desde la matriz epistémica asumida por los autores, aplicándose las normas APA, Sexta Edición, proceso de anti plagio en línea Plagiarisma, garantizándose así la cientificidad de la obra.

Comité Editorial

ÍNDICE GENERAL

PRÓLOGO

Los Modelos Lineales fueron utilizados a lo largo de décadas tanto intensa como extensivamente en aplicaciones Estadísticas. Llamamos Modelos Lineales a esas situaciones que luego de haber sido analizadas Matemáticamente, se representan mediante una funcionalidad lineal, los cuales son lineales en las fronteras desconocidas e integran un elemento de error. El elemento de error es el que los convierte en Modelos Estadísticos. Dichos modelos son la base de la metodología que habitualmente llamamos Regresión Múltiple. Por esta razón el funcionamiento de los Modelos Lineales es imprescindible para entender y ejercer correctamente los Procedimientos Estadísticos.

En algunas ocasiones el modelo coincide claramente con una recta; en otras ocasiones, a pesar de que las variantes que interesan no pertenecen cada una de a la misma línea, es viable hallar una funcionalidad lineal que mejor se aproxime al problema, ayudando a obtener información preciada.

Un Modelo Lineal se puede establecer de forma gráfica o bien, mediante una ecuación. Hay situaciones en que en una de las variantes se desea que cumpla numerosas condiciones a la vez, entonces nace un grupo de ecuaciones donde el punto de intersección de dichas ecuaciones representa la solución del problema.

La Modelación necesita precisamente de supuestos, puesto que de otra forma no podríamos representar a escala y con sencillez una realidad compleja. Un óptimo modelo podría ser ese que se enfoque primordialmente en explicar la verdad, empero además ese que tenga capacidad de hacernos ver más allí de lo cual a primera vista parece dar. Un modelo “malo” es ese enormemente realista, empero tan difícil que se vuelve inmanejable; en esta situación no hay razón para construirlo.

INTRODUCCIÓN

El Análisis Estadístico de las relaciones individuales que forman un modelo, y del modelo como un conjunto, hace posible adjuntar una medida de confianza a los pronósticos del modelo. Una vez que se ha construido un modelo y se ha adecuado a los datos, puede usarse un análisis de sensibilidad para estudiar muchas de sus propiedades. En particular, pueden evaluarse los efectos de cambios pequeños en variables individuales en el modelo. Por ejemplo, en el caso de un modelo que describe y predice tasas de interés, uno podría medir el efecto en una tasa de interés particular de un cambio en el índice de inflación, este tipo de estudio de sensibilidad sólo puede realizarse si el modelo está en forma explícita.

Comúnmente se aplican o se realizan pronósticos de una manera u otra. Pocos reconocen, no obstante, que alguna clase de composición lógica o modelo, está implícita en cada predicción. Por consiguiente, inclusive un pronosticador intuitivo construye cualquier tipo de modelo, quizá sin darse cuenta de que lo hace. Edificar modelos impone al sujeto a pensar con claridad y describir cada una de las interacciones relevantes implicadas en un problema.

Fiarse de la intuición podría ser inseguro en ocasiones gracias a la probabilidad de que se ignoren o se utilicen de forma inapropiada interrelaciones relevantes. Además, es fundamental que las colaboraciones particulares sean validadas de alguna forma. Empero, principalmente no se hace esto una vez que se hacen pronósticos intuitivos. No obstante, en el proceso de edificar un modelo, una persona debería validar no únicamente el modelo en general sino además las interrelaciones particulares que conforman el modelo.

Se desarrollará la teoría de los Modelos de Regresión Lineal Sencilla, Estimación y Prueba de Premisa, Validación del Modelo y Pronóstico, Modelos de Regresión Lineal Múltiple, Pruebas de los Límites y Validación del Modelo de Regresión Lineal Múltiple,

Modelos de Regresión con Variantes Cualitativas, otros Modelos e Inconvenientes, y Procedimientos de Selección de Variantes.

Para el desarrollo de los ejemplos o aplicaciones que se realizaran se va a hacer uso del paquete estadístico SPSS v15.0. En todos los capítulos se muestra una pequeña introducción, así como además una definición de términos básicos. Al final se muestran los apéndices y las referencias bibliográficas que se han utilizado a lo largo de la indagación.

RESUMEN

El análisis de regresión lineal es una técnica estadística usada para estudiar la relación entre variables. Se adapta a una amplia variedad de situaciones y se enmarca en modelos lineales. Su objetivo es hallar una relación que describa o resuma la relación entre dos o más variables; por ejemplo: ¿cómo varía el promedio anual del maíz, según producción a nivel nacional? o ¿cómo varía el consumo de gasolina de un auto, según peso y potencia de su motor? En contexto de la investigación de mercados puede utilizarse para determinar en cuál de diferentes medios de comunicación puede resultar más eficaz invertir o predecir el número de ventas de un determinado producto. En caso de dos, regresión simple, o más variables, regresión múltiple, el análisis de regresión lineal puede usarse para explorar y cuantificar la relación entre una variable llamada dependiente, endógena y una o más variables independientes, exógenas, así como desarrollar una ecuación lineal predictiva. Asocia una serie de procedimientos de diagnóstico, análisis de residuos y puntos de influencia respecto a estabilidad e idoneidad de análisis. Probablemente, es la técnica más usada para establecer relaciones funcionales entre variables a partir de una relación lineal de forma $Y = X\beta + U$.

Actualmente, es una herramienta metodológica cuantitativa y cualitativa que permite la asignación óptima de recursos escasos, apoyar eficientemente el proceso de toma de decisiones respecto a maximizar ganancias, utilidades, satisfacción de clientes o minimizar costos, distancias y tiempos. Hace uso de modelos matemáticos, que procesan datos, representan el problema real, establecen hipótesis que es una representación precisa de la realidad tal que sus conclusiones sean válidas para resolver un problema práctico real; por ejemplo: cuánto producir con base en variables estocásticas-no estocásticas, dónde producir, a qué precio ofertar, a qué precio comprar materia prima, medios y rutas optimas de transporte.

Con base en esto, Regresión Lineal con Ofimática presenta terminología básica de regresión lineal, demostraciones matemáticas superiores, modelado matemático, soluciones factibles, cálculos iterativos de maximización, minimización con método algebraico con método de ecuaciones simultáneas, programación lineal, algoritmos matemáticos, entre otras. Con información del Censo Nacional Agropecuario (CNA, 2000), Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC, 2002 a 2020) del Instituto Nacional de Estadística y Censos (INEC) del Ecuador a nivel Nacional, Regional, Provincial y Cantonal de cultivos permanentes (banano, cacao, café, caña de azúcar para biocombustible, caña de azúcar tallo fresco, naranja, palma africana, plátano y tomate de árbol), transitorios (arroz, cebada, fréjol, maíz duro seco, maíz suave choclo, papa, trigo y yuca), producción animal (asnal, caballar, caprino, mular, pollos-gallinas en campo, pollos-gallinas en planteles avícolas, porcino y ovino), pastos (cultivados y naturales) y montes-bosques mediante uso de software con licencia (Microsoft Excel), open office (Python-Jupyter, R y R-Studio) e interactúa amplia y profundamente con Estadística para ingeniería, Diseño experimental para ingeniería e Investigación operativa para ingeniería, que paralelamente usan software con licencia y open office mencionados para desarrollo de ejercicios teóricos-prácticos complementarios.

CAPÍTULO I

CAPÍTULO I

MÁXIMA VEROSIMILITUD (**MV**)

1.1. Predicción

Uno de los objetivos de regresión es predicción y pronóstico de nuevos valores de variable dependiente. Dado un valor $x_0 \notin \{x_1, x_2, x_3, x_4, \dots, x_n\}$ sería deseable predecir o adivinar el valor correspondiente $y_0 \in Y$.

1.1.1. Consideraciones generales

El problema de predicción es conceptualmente diferente del problema estimación. En estimación el objetivo es dado un parámetro θ no conocido, pero fijo y una muestra genérica $\{Y_1, Y_2, Y_3, Y_4, \dots, Y_n\}$ hallar una función de muestra $\theta\{Y_1, Y_2, Y_3, Y_4, \dots, Y_n\}$ que, en promedio, sobre todas las muestras posibles, esté cerca de θ:

$$E(\theta\{Y_1, Y_2, Y_3, Y_4, \dots, Y_n\} - \theta)^2 \text{ minimiza sobre } \hat{\theta}$$

Esta esperanza es error cuadrático medio $ECM_{(\theta)}(\hat{\theta})$. En predicción se tiene una muestra concreta $\{y_1, y_2, y_3, y_4, \dots, y_n\}$ y su objetivo de predecir una realización de variable aleatoria Y. Se nombra $\hat{y}_{0n}$ el predictor de y tal que el error de predicción es:

$$e_{(0|n)} = y - \hat{y}_{(0|n)}$$

Un criterio muy aceptado para determinar $\hat{y}_{(0|n)}$ es minimizar el error cuadrático medio de $\hat{y}_{(0|n)}$:

$$ECM_{\left(\hat{y}_{(0|n)}\right)} = \min_{\hat{y}_{(0|n)}} E\left(y - \hat{y}_{(0|n)}\right)^2$$

1.1.2. Predicción con modelo regresivo

Si tiene un modelo de regresión válido, apropiado a datos:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2(x)$$

Tal que, es "natural" pronosticar y_0 mediante:

$$\hat{y}_{(0|n)} = \hat{\beta}_1 + \hat{\beta}_2(x_0)$$

Tal que $\hat{\beta}_1, \hat{\beta}_2$ son estimadores de mínimos cuadrados de $\hat{\beta}_1$ y $\hat{\beta}_2$. Sin embargo, el predictor y variable pronosticada tienen esperanza igual:

$$E\left[\hat{y}_{(0|n)}\right] = E[\hat{\beta}_1] + E[\hat{\beta}_2](x_0) = \beta_1 + \beta_2(x_0) = E(y_0)$$

Tal que, varianza del predictor $\hat{y}_{(0|n)}$ es:

$$e_{(0|n)} = [\beta_1 + \beta_2(x_0) + u_0] - (\hat{\beta}_1 + \hat{\beta}_2[x_0]) = u_0 - (\hat{\beta}_1 - \beta_1) - (\hat{\beta}_2 - \beta_2)[x_0]$$

Donde, por independencia de u_0 respecto a $\{\hat{\beta}_1, \hat{\beta}_2\}$ error cuadrático medio de predicción, varianza $e_{(0|n)}$, es:

$$\begin{aligned} \underset{\text{Error cuadrático medio}\left(\hat{y}_{(0|n)}\right)}{ECM} &= E\left[u_0 - (\hat{\beta}_1 - \beta_1) - (\hat{\beta}_2 - \beta_2)[x_0]\right]^2 \\ &= E(u_0^2) + E\left[(\hat{\beta}_1 - \beta_1) - (\hat{\beta}_2 - \beta_2)[x_0]\right]^2 \\ &= E(u_0^2) + Var\left[(\hat{\beta}_1 + \hat{\beta}_2 x_0) - (\beta_1 - \beta_2 x_0)\right] = E(u_0^2) + Var\left[(\hat{\beta}_1 + \hat{\beta}_2 x_0)\right] \end{aligned}$$

Por igualdad $Var[\hat{y}_i] = \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$:

$$\underset{\text{Error cuadrático medio}\left(\hat{y}_{(0|n)}\right)}{ECM} = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

Si el objetivo es predecir media de y_0; es decir $E[\hat{y}_0] = \beta_1 + \beta_2 x_0$, la predicción $\hat{y}_{(0|n)}$ no cambia más si es el mejor predictor lineal de media, pero si cambia el error:

$$e_{(0|n)} = [\beta_1 + \beta_2(x_0)] - (\hat{\beta}_1 + \hat{\beta}_2[x_0]) - (\hat{\beta}_1 - \beta_1) - (\hat{\beta}_2 - \beta_2)[x_0]$$

Error cuadrático media es:

$$\underset{\text{Error cuadrático medio}}{\text{ECM}}\left(\hat{y}_{(0|n)}\right) = \text{Var}\left[\hat{y}_{(0|n)}\right] = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

1.2. Intervalos y bandas de confiabilidad

Si estima σ^2 mediante $S_R^2 = \left(\frac{\text{SRC}_{\text{(Suma de Residuos al Cuadrado)}}}{(n-2)}\right)$, entonces:

$$\frac{\hat{y}_{(0|n)} - y_0}{\sqrt{S_R^2\left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]}} \rightarrow t_{(n-2)}$$

$$\frac{\hat{y}_{(0|n)} - E(y_0)}{\sqrt{S_R^2\left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]}} \rightarrow t_{(n-2)}$$

Donde el intervalo de confianza de nivel $1 - \alpha$ para predecir y_0 es:

$$\hat{y}_{(0|n)} \pm \sqrt{S_R^2\left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]}\, t_{(n-2)}\left(\frac{\alpha}{2}\right)$$

Se dice que este es intervalo de confianza alrededor de recta de regresión, la unión continua de extremos de estos intervalos forma una hipérbola entorno de recta de regresión al igual que extremos de intervalos de confianza para predicción. Es importante considerar que longitud mínima de intervalos de confianza se alcanza en $x_0 = \bar{x}$ y crece con valor absoluto de su diferencia. Si x_0 se aleja mucho de $\bar{x}$ el intervalo de confianza será muy impreciso, si se une el hecho que el modelo estimado puede tener validez únicamente en rango observado de x, se concluye no debe extrapolar "muy lejos" del rango.

1.3. Transformaciones de modelos no lineales a lineales

En muchas ocasiones los modelos no lineales pueden ser tratados como lineales si se realizan algunas transformaciones a las variables, ya sea a la predictora, a la respuesta o a ambas. Al emplear tales transformaciones se deberá tener la precaución de verificar que el modelo modificado cumple con la hipótesis sobre la distribución que siguen los errores.

1. **Modelo exponencial:**

$$y = e^{(\beta_0+\beta_1+\varepsilon)}.$$

Tomando logaritmos en ambos miembros:

$$\text{Ln}\,(y) = \beta_0 + \beta_1 + \varepsilon$$

Si ponemos $z = \text{Ln}\, y$ queda $z = \beta_0 + \beta_1 x + \varepsilon$, es un modelo lineal simple que se estima con $\hat{z} = b_0 b_1 x$.

Modelo recíproco o inverso:

u modelo es:

$$y = \frac{1}{\beta_0+\beta_1 x + \varepsilon}$$

Con ingresos en ambos miembros:

$$\frac{1}{y} = \beta_0+\beta_1 x + \varepsilon$$

Si $z = \frac{1}{y}$ en última igualdad es:

$$z = \beta_0+\beta_1 x + \varepsilon$$

Que es un modelo lineal usual.

2. **Modelo multiplicativo o potencial:**

El modelo multiplicativo, también conocido como potencial, es:

$$y = \alpha x^{\lambda} \varepsilon$$

Tomando logaritmos en ambos miembros de la igualdad:

$$\text{Ln } (y) = \text{Ln}(\alpha) + \lambda \, \text{Ln}(x) + \text{Ln}(\varepsilon)$$

Haciendo $z = \ln y, t = \ln x, \beta_0 = \ln \alpha$ y $\beta_1 = \lambda$, el modelo se escribe:

$$z = \beta_0 + \beta_1 t + \varepsilon$$

Que se estima por el modelo lineal $\hat{z} = b_0 + b_1 t$.

Otros modelos no lineales comúnmente utilizados son:

3. Logarítmico. $y = \alpha + \beta \text{Ln}(x)$, se utiliza la transformación $t = \text{Ln}(x)$.
4. Compuesto. $y = \alpha \beta^x$, se linealiza mediante $\text{Ln}(y) = \text{Ln}(\alpha) + \text{Ln}(\beta x)$.
5. Curva S. $y = e^{(\alpha + \beta/x)}$, que se transforma en $\text{Ln}(y) = \alpha + \beta t$ con $t = \frac{1}{x}$.

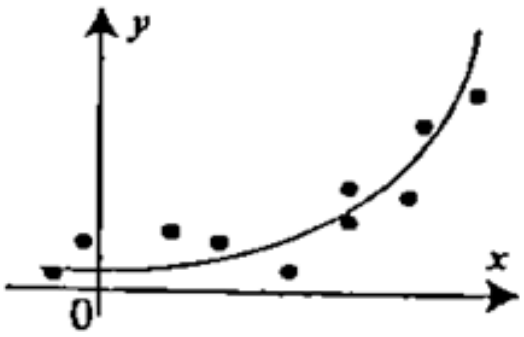

Figura 1. *Modelos no lineales comúnmente utilizados*

Se recomienda que el lector realice las operaciones necesarias para linealizar estos modelos.

1.4. Ejemplos

1.4.1. Velocidad en función de densidad

Se conoce que entre más vehículos circulan en una carretera, menos es la velocidad del flujo de tráfico. Estudiar la relación entre estas variables tiene como objetivo la mejora del transporte y reducción de tiempo de viaje.

1.4.2. Búsqueda de modelo

Para construir un modelo se requiere información o datos, con este propósito se observará o experimentará. Por ejemplo: se puede tomar una carretera cualquier, contar número de vehículos en una determinada extensión y velocidad promedio en que se transportan.

1.4.3. Uso de modelo predictivo

Se usan puntos de alta densidad para hallar un modelo adecuado para esos puntos y, después, se emplea el modelo para estimar velocidad de puntos de baja densidad. Una vez que el modelo es satisfactorio habrá que emplearlo para realizar estimaciones y predicciones que servirán para analizar el comportamiento de la variable repuesta ante condiciones que no fueron probadas empíricamente. Previamente se establece la diferencia entre intervalos de estimación y de predicción:

1. Un intervalo de estimación se utiliza para estimar la media de Y, dado un valor de x; es decir estimar el valor $\mu_{(Y|x_0)}$. Un investigador puede desear conocer el consumo medio, para un peso dado del auto.
2. Un intervalo de predicción se utiliza para estimar un valor particular de Y, cuando x = x; es decir, estimar el valor de Y_0 mediante un intervalo de confianza. El investigador puede desear conocer el consumo d un auto que tiene un peso dado.

Nótese que los valores de estimación y de predicción de Y son idénticos en los dos casos, la diferencia radica en la precisión relativa de cada una, que se ven reflejadas en sus varianzas e intervalos de confianza.

Cuyo resultado es el sistema de ecuaciones:

$$X^tXb = X^tY$$

Las matrices $X^tXb = X^tY$ son:

$$X^tX = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i \end{pmatrix} \quad y \quad X^tY = \begin{pmatrix} \sum_{i=1}^{n} Yi \\ \sum_{i=1}^{n} x_i\ y_i \end{pmatrix}$$

Si la matriz X^tX es inversible, se llega a la ecuación de estimación de los parámetros:

$$b = (X^tX)^{-1}X^tY$$

La matriz $(X^tX)^{-1}$ es:

$$(X^tX)^{-1} = \frac{1}{n\sum^{n} x_i^2 - \left(\sum_{i=1}^{n} x_1\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i^2 & n \end{pmatrix}$$

Con esto, los valores ajustados $\widehat{Y}$ se obtiene evaluando:

$$\widehat{Y} = Xb.$$

Y la ecuación de predicción de un valor y_p correspondiente a un valor x_p dado es:

$$y_p = x_p^t b_1$$

$$\text{Con } x_p^t = (1; x_p).$$

1.5. Servicios Eco sistémicos Hídricos (SEH) Forestales en La Delegación La Magdalena Contreras, Distrito Federal, México

Este estudio se realizó con 230 encuestas a usuarios del Servicio Ecosistémico Hídrico, Delegación La Magdalena Contreras, Distrito Federal, que abastece a 65 408 hogares. Se diseñó un modelo de ecuaciones simultáneas para analizar su disposición a pagar. Los resultados estadísticos indican multicolinealidad en estos modelos, donde las variables independientes y regresores, son linealmente independientes.

En consecuencia, el hecho que esté cada una relacionada con la variable dependiente implica que, de alguna manera, estén a su vez relacionadas entre sí. Algunas soluciones son aumentar la información, introduciendo nuevas variables que amplíen la explicación del modelo, pues la multicolinealidad es un problema de falta de información (Intriligator, 1978, citado por (Ramírez, 2001)).

Otra solución es reducir el número de variables explícitas, aquellas que no aporten información adicional, aunque se puede caer en error de la especificación (Myiers, 1986, citado por (Ramírez, 2001)). Asimismo, la posibilidad de utilizar el modelo es otra solución, pero es riesgoso hacerlo si se trata de explicar la estructura que el modelo representa.

Además, la multicolinealidad alta ni siquiera permite hacer buenas predicciones. Finalmente, existen procedimientos de estimación opcionales para corregirla, que se han agrupado bajo el nombre de Técnicas de Estimación Sesgadas (Ramírez, 2001).

Los objetivos particulares son:

1. Estimar la disposición a pagar de usuarios consuntivos domésticos del agua para mejorar el servicio ecosistémico hídrico en el área de estudio;

2. Obtener un modelo estadísticamente significativo, cuyas variables independientes expliquen la disposición a pagar;

3. Evaluar la calidad del agua de la red pública de suministro y el agua purificada en garrafones de tres principales marcas comerciales que se venden y

4. Obtener el costo económico defensivo de la compra de agua purificada y conocer el costo por asegurar agua potable mediante la red pública de suministro.

El objetivo general es articular conocimientos interdisciplinarios de ciencias de políticas para instituir una política hídrica pública incentivo-distributiva en este Municipio.

Las variables empleadas para medir la valoración del servicio ecosistémico hídrico de los usuarios consuntivos domésticos fueron: consumo de agua, cantidad de garrafones de 20 l/mes que compra, marcas comerciales que prefiere el usuario, consumo de agua en l/semana para uso doméstico, pago económico anual por servicio público de agua actual y en años anteriores, fuente de suministro de agua potable, frecuencia de servicio público, calidad de agua, opinión del usuario si el dinero que paga por el servicio público es el adecuado y si cuenta con algún sistema de almacenamiento de agua; monto económico máximo que el usuario consuntivo doméstico está dispuesto a pagar por mejorar el servicio ecosistémico hídrico en categorías monetarias, monto anual de su cooperación, mecanismo de cobro, instancia apropiada que administre el dinero, instrumento de recaudación más apropiado y motivo por el que no cooperaría para mejorar el servicio ecosistémico hídrico y perfil socioeconómico de los encuestados, número de personas que conforman su familia, sexo, estado civil, edad, nivel de estudios, ocupación y nivel de ingresos totales anuales por familia.

Las variables a medir en el modelo econométrico de mínimos cuadrados ordinarios son: consumo doméstico de agua potable medido en cantidad de litros de agua consumidos por año (CD_i), pago anual por consumo de agua medido en cantidad

monetaria (PCA_i), disposición a pagar por el servicio ecosistémico hídrico medido en cantidad monetaria anual a cooperar (DAP_i), ingreso familiar total anual medido en términos monetarios (I_i), nivel académico de estudios del encuestado medido en años académicos (E_i), cantidad de agua potable depositada en los sistemas de almacenamiento del usuario entrevistado medida en litros de agua almacenados (QAA_i), consideración del encuestado si el pago por el servicio público de agua es adecuado medido en categoría dicotómica o variable muda (PSA_i) y sexo del encuestado medido en categoría dicotómica (S_i).

De estos cuestionarios, se seleccionaron las variables dependiente disposición a pagar por mejorar el servicio ecosistémico hídrico en términos monetarios anuales (Y) y variables independientes ingreso familiar total en términos monetarios anuales (X_1), nivel académico medido en años escolares (X_2) y pago monetario anual promedio por servicio público de agua potable, en el periodo de 2005 a 2008 (X_3) para hacer el modelo estadístico de logaritmos naturales.

1.6. Encuesta de superficie y producción agropecuaria continua (ESPAC Provincial) 2000-2020, Ecuador

La información agropecuaria del Ecuador se obtiene a través de la Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC), este proyecto viene ejecutándose desde el año 2002 hasta la actualidad; desde su primer diseño y construcción de los marcos muestrales, no ha existido una actualización de los mismos.

La desactualización de los marcos muestrales, ha provocado que la estimación de los resultados pierda representatividad de áreas geográficas y productos, además no permitía obtener información desagregada por provincia para la región oriental. Bajo estos antecedentes, el INEC decidió realizar la actualización del marco muestral, lo que permite mejorar los niveles de estimación de los resultados agropecuarios provistos por la ESPAC, con la finalidad principal de obtener y

producir datos que midan de manera permanente la dinámica del sector agropecuario de forma científica, moderna, eficiente y con innovación tecnológica.

Para la ESPAC 2014, el INEC realizó una actualización del marco maestro de muestreo utilizado para la encuesta. Esta actualización implicó la utilización de información cartográfica digital, la cual sirvió de insumo para generar una nueva estratificación. En el año 2016 se realizó la investigación de 5.638 segmentos. Los resultados tienen representatividad a nivel provincial. Se trata de un Marco Múltiple de Muestreo (MMM), que combina marco de áreas (MM) y marco de lista (ML).

El material cartográfico de base utilizado para la construcción del marco es el Mapa de Usos del Suelo elaborado por el MAGAP y actualizado durante el período 2012-2014. Este mapa cubre el 85% del territorio nacional, aunque a diferentes escalas: 1:25000, 1:50000 y 1:100000. El 15% restante se cubre con imágenes de satélite.

El territorio nacional se estratifica en dos estratos primarios: (i) las zonas urbanas densamente pobladas y (ii) el resto del territorio, incluyendo áreas agropecuarias, forestales, agua y otras áreas rurales con baja densidad de población. Del estrato (ii) se separan solo los cuerpos de agua. El resto (áreas agropecuarias y forestales, reservas naturales y otras áreas no agropecuarias) se estratifican en cuatro estratos, según la proporción de superficie cultivada.

Los límites entre estratos son los estándares recomendados en FAO (1996,1998)]: áreas con una proporción mayor del 60% (Estrato1 y 2); áreas con una proporción de superficie cultivada entre 20% y 60% (Estrato 3) y áreas con una proporción de superficie cultivada inferior al 20% (Estrato 4). El Estrato 1, se divide en dos estratos, atendiendo al tamaño de los terrenos (parcelas o campos): terrenos pequeños (Estrato 1) y terrenos grandes (Estrato 2).

En primer lugar, se segmenta todo el territorio en una retícula de celdillas cuadradas de 576 hectáreas. Esta segmentación se realiza sobre el mapa de usos y se calcula el

porcentaje de tierras cultivadas (incluyendo las praderas cultivadas) en cada celdilla. Las celdillas se clasifican en los estratos 1 al 4, en función de ese porcentaje.

Las celdillas del Estrato 1 se clasifican a su vez en dos subestratos: Sierra (Estrato 01) y el resto (Estrato 02). Se consideran segmentos de límites geométricos. El criterio usual para elegir el tamaño del segmento consiste en fijar el número idóneo de unidades de observación dentro de sus límites y ese número idóneo se ha fijado en 10 terrenos.

A partir del marco de áreas se selecciona una muestra aleatoria de segmentos. Sobre los segmentos de la muestra se observa parte de la información requerida (la superficie de los cultivos y demás usos del suelo y algunas variables ambientales, tales como tipos de riego o técnicas de laboreo). Mediante entrevistas a los agricultores que cultivan las tierras dentro de los límites del segmento, se recoge toda la información de las variables que contiene el cuestionario ESPAC.

Las tasas de muestreo consideradas son, respectivamente, del 3%, 2%, 1% y 0,5% para los estratos 1 (Superficie cultivada>60% y campos pequeños (Sierra)), 2 (Superficie cultivada>60% y campos grandes), 3 (20%< Superficie cultivada<60%) y 4 (Superficie cultivada<20%). Se espera que esos tamaños de muestra sean suficientes para que las precisiones de las estimaciones de los cultivos mayoritarios estén dentro de los límites de tolerancia estándares (15% de coeficiente de variación).

Sin embargo, es de temer que la precisión de las estimaciones de los cultivos muy localizados en el espacio (como los hortícolas, y algunos cultivos industriales) y la ganadería, supere los límites de tolerancia: para mejorar esas precisiones se usarán marcos múltiples, complementando la muestra de áreas con una lista de los mayores productores. Para la ESPAC 2014, se seleccionó un total de 5,638 segmentos en el marco de áreas y para el marco de lista se seleccionaron 3,969 unidades productivas, distribuidas en 14 estratos:

1.6.1. Arroz (Oriza sativa)

Es una especie perteneciente a familia Poáceas (gramíneas), su semilla es comestible y constituye la base de la dieta de casi la mitad de la población mundial. Su nutriente principal son los hidratos de carbono; aunque, aporta proteínas, 7 %, minerales y, natural, vitaminas. Es el cultivo de ciclo corto más extensamente cultivado en Ecuador. Según la Encuesta de Superficie y Producción Agropecuaria Continua – SPAC - 2011, 378.643 hectáreas fueron dedicadas a este cultivo.

En términos nutricionales, esta gramínea es la que mayor aporte de calorías brinda a la dieta de todos los ecuatorianos (FAO). USDA Post pronostica que la producción de arroz elaborado de Ecuador en 2019-20 (abril-marzo) será de alrededor de 870,000 toneladas, frente a 925,000 toneladas estimadas para 2018-19. La disminución se atribuye a una caída en el área de siembra y a rendimientos ligeramente más bajos. Incluso, esta disminución en la superficie de siembra de arroz se debe a que "los agricultores más pequeños y más antiguos se retiran, pues no pueden competir con los agricultores más productivos y grandes que usan mejores prácticas agronómicas".

1.6.2. Banano (Musa paradisiaca)

Sus nombres comunes son banano, banana, plátano, cambur, topocho, maduro y guineo. Hacen referencia a un gran número de plantas herbáceas del género Musa y tanto híbridos obtenidos horticulturalmente, a partir de especies silvestres ***Musa acuminata*** y ***Musa balbisiana***, como cultivares genéticamente puros de estas especies se aprovechan. Según estadísticas, se conoce que el Ecuador se inició en la exportación de banano en el año 1910, exportando 71,617 racimos de más de 100 libras.

El Estado Ecuatoriano ha intervenido en la actividad bananera desde que inicia el cultivo en gran escala. En nuestro país la verdadera comercialización bananera se

inicia en la década de 1950; aunque, en la Provincia de El Oro se tiene registro de su producción desde 1925 comercializando hacia los mercados de Perú y Chile.

El desarrollo de la actividad bananera ha estado muy vinculada a iniciativa privada de ecuatorianos que han invertido su capital tanto económico como humano en actividades de producción y exportación de la fruta. Ha recibido la valiosa contribución de capitales internacional que ha permitido que el Ecuador sea el primer país exportador de banano en el mundo con aproximadamente un 30% de la oferta mundial, seguidos por Costa Rica, Filipinas y Colombia, juntos abastecen más del 50% del banano consumido en el mundo.

1.6.3. Cacao ramilla (Theobroma cacao L)

Es una pequeña planta tropical que se cultiva por sus semillas en forma de almendra, usadas para elaborar el chocolate. Es llamado árbol del cacao o cacaotero. Pertenece a la familia de las malváceas. La especie es originaria del bosque tropical de la cuenca del Amazonas y se reconocen dos zonas de distribución en era precolombina. Se cultivó por primera vez en Centroamérica, norte de Suramérica y las variedades que allí se encontraron se conocen como criollas.

El Ecuador exporta cacao en 3 diferentes formas, que se refieren a etapas distintas de elaboración: granos de cacao, semi-elaborados y producto terminado. Por sus condiciones geográficas y su riqueza en recursos biológicos, Ecuador es el productor por excelencia de Cacao Arriba fino y de aroma, pues 63% de la producción mundial proveniente de la variedad Nacional.

Este tipo de grano es utilizado en todos los chocolates refinados. Sin embargo, el chocolate fino se distingue por su pureza; específicamente, el sabor y fragancia que el cacao tiene. Del total de la exportación ecuatoriana se estima que un 75% es cacao fino de aroma mientras que el restante 25% pertenece a otras variedades como el CCN51. Ecuador se posiciona como el país más competitivo de América Latina en este campo, seguido de lejos por Venezuela, Panamá y México, que son países que

poco a poco han incrementado su participación en el mercado mundial del cacao fino en grano.

1.6.4. Café robusta fresca o madura (Coffea canephora)

Modalidad de café, tiene el grano de café más pequeño, es resistente y posee un mayor nivel de cafeína. Se cultiva en África, Asia y Brasil. Fue descubierto en Africa a fines del siglo XIX, pues crece de manera silvestre en zonas tropicales de El Congo y Guinea. Entre los años 1951 y 1986 se realizaron introducciones de germoplasma de café robusta hacia el Ecuador, desde el centro Agronómico Tropical de Investigación y Enseñanza (CATIE-Costa Rica).

Las introducciones de café robusta se establecieron en bancos de germoplasma de la Estación Experimental Tropical Pichilingue del INIAP, ubicada en la Provincia de los Ríos. Café robusto se fue dispersando progresivamente, desde la estación Pichilingue hacia otras zonas cercanas, específicamente en cantones Quevedo, Mocache, Ventanas y otros. En el año 1968, debido a la migración de agricultores hacia la Amazonia se produjo la diseminación de café robusta hacia esas localidades. Cabe indicar que la propagación del café robusta, hasta 1990, se realizaba solo por la vía sexual; es decir, usando plantas provenientes de semilla, que debido a su naturaleza alogamica, generó alta variabilidad fenotípica en cafetales.

1.6.5. Caña azúcar tallo fresco (Saccharum officinarum)

Son especies de herbáceas, vivaces, de tallo leñoso de un género (***Saccharum***) de la familia de gramíneas (Gramineae), originaria de la Melanesia y cuya especie fundamental es ***Saccharum officinarum***. Fue introducida en Cuba por el año 1535 desde Santo Domingo. La caña de azúcar se cultiva mucho en países tropicales y subtropicales de todo el mundo por el azúcar que contiene en los tallos, formados por numerosos nudos.

1.6.6. Caña azúcar biocombustible (Saccharum officinarum)

Una característica de la caña de azúcar como fuente de biocombustibles es su capacidad para generar un sinfín de productos. Mientras que en España se cultiva para producir azúcar y etanol, países líderes en su producción, como Brasil o India, obtienen también electricidad con la combustión del bagazo, vinaza y compuestos químicos como furfural y amoniaco. También, se produce biodiesel a partir de los azúcares de la caña.

Según informa Al Costa, director general de Alcohol, consultora española especializada en etanol y caña de azúcar, en Brasil funcionan cientos de autobuses con biodiésel a partir de caña de azúcar, obtenido gracias a la modificación genética de la levadura de la cerveza. La modificación consigue que el microorganismo secrete una molécula llamada farneseno, común en el diésel e incluso en muchas plantas (responsable del olor en algunas manzanas). A partir de aquí, puede utilizarse en cualquier motor diésel sin necesidad de ajustes.

La gran ventaja de esta tecnología es que se basa en la fermentación, un método conocido y usado en todas las plantas de etanol hasta las de celulósico; por lo que, la aparición de sorpresas indeseables en procesos es nula. Esto permite el uso de cualquier planta de etanol también para la obtención del biodiésel, pues en lugar de usar las cepas normales de ***Saccharomyces*** (u otras) para fabricar el primero, se emplean cepas genéticamente modificadas para producir farneseno, separado vía destilación normal del etanol y elaborar biodiésel. El efecto final es la producción de ambos biocarburantes en un único proceso, con los consiguientes beneficios económicos.

1.6.7. Cebada (Hordeum vulgare)

La cebada es un alimento de origen vegetal perteneciente a clasificación de cereales. Se trata de una planta gramínea y crece de la misma manera que otros cereales como el trigo. Se caracteriza por ser un cultivo denominado como el cereal en invierno,

pues su cosecha se lleva a cabo en verano y la distribución de Asia de manera similar el trigo en la época fría del año.

La cebada es altamente conocida por ser un cereal que se emplea en la industria cervecera y es un alimento reconocido por la cantidad de nutrientes qué aporta al organismo. Ochenta hectáreas de cebada -30 de la variedad Scarlett y 50 de Cañicapa- se siembran en cinco cantones de Tungurahua, a través del Programa Siembra Cebada, que impulsan el Ministerio de Agricultura, Ganadería, Acuacultura y Pesca (MAGAP), y la Cervecería Nacional.

Las dos entidades fomentan la producción de cebada maltera, con la finalidad de elevar la productividad y los ingresos de los productores. José Luis Broncano, técnico de Cervecería Nacional, indicó que la aplicación de este programa es una prioridad, pues la empresa importa más de 30 mil toneladas anuales para elaborar la cerveza. En Tungurahua, el Programa Siembra Cebada empezó en noviembre de 2015. El precio del quintal es de 25 dólares.

1.6.8. Frejol seco (Phaseolus vulgaris)

Es la especie más conocida del género Phaseolus en familia Fabaceae. Sus semillas y su planta reciben en el mundo hispanohablante diversos nombres según el país o región, pero los más comunes son: frejol, fríjol, frijol, habichuela, caraota, poroto, judía y alubia. El frijol es una especie anual nativa de Mesoamérica y sus numerosas variedades se cultivan en todo el mundo para el consumo, tanto de sus vainas verdes como de sus semillas frescas o secas. Fue introducido en América por las tribus nómadas que cruzaron el estrecho de Bering hasta Alaska.

Hay evidencias que en siglo X los Aztecas en México usaron el fríjol como un grano básico e Incas lo introdujeron a Suramérica. Es la leguminosa de grano de consumo humano directo más importante en el planeta; ocupa el octavo lugar entre las leguminosas sembradas en el mundo. Para la población ecuatoriana constituye una de las principales fuentes de proteína y carbohidratos.

Mundialmente el frijol es la leguminosa alimenticia más importante para cerca de 300 millones de personas, que, en su mayoría, viven en países en desarrollo, debido a que este cultivo, conocido también como "la carne de los pobres", es un alimento poco costoso para consumidores de bajos recursos. El frijol es especialmente importante en la alimentación de mujeres y niños; además, tiene gran importancia económica, pues genera ingresos para millones de pequeños agricultores.

A nivel mundial se producen 18.991,954 t, siendo los mayores productores mundiales: Brasil (3 millones de t), India (2.9 millones de t), México (1.5 millones de t) Nicaragua, Myanmar (1.9 millones t), China (1.9 millones t) entre otros países. Ecuador produce 39,725 t, es decir, el 0.2% de la producción mundial.

1.6.9. Maíz duro seco (Zea mays L)

Es una gramínea anual originaria y domesticada por los pueblos indígenas en el centro de México desde hace unos 10 000 años e introducida en Europa en el siglo XVII. Los indígenas taínos del Caribe denominaban a esta planta mahís, que significa literalmente "lo que sustenta la vida".

Actualmente, es el cereal con el mayor volumen de producción a nivel mundial, superando incluso al trigo y al arroz. El maíz duro seco es importante para la industria de balanceados y productores de aves. Según Víctor Romero, presidente de la Federación de Avicultores del Ecuador, los pequeños productores están a punto de la quiebra por el abaratamiento de la carne en las granjas. Esto por la sobreproducción e importación de huevos fértiles.

En la sierra del Ecuador el cultivo de maíz (***Zea Mays*** L.) es uno de los más importantes debido a la superficie sembrada, el papel que cumple en la seguridad y soberanía alimentaria, al ser un componente básico de la dieta de la población rural. La superficie sembrada de maíz en las provincias de la sierra ecuatoriana para el año 2011 fue de 168486 Ha y el consumo per cápita de maíz es alrededor de 14,50 Kg/año. En el país, el año pasado en invierno se sembraron 276.385 hectáreas (ha)

de maíz duro seco. Según Ministerio de Agricultura, Ganadería, Acuacultura y Pesca (MAGAP), Los Ríos lleva la delantera en producción (35,2 %), seguida de Manabí (28,9 %) y Guayas (17,9); el resto se reparten entre Loja, Santa Elena y El Oro.

1.6.10. Maíz suave choclo (Zea mays amylacea)

Sus variedades tradicionales constituyen un rico patrimonio de tradiciones agrícolas y alimenticias. En el Ecuador el maíz se cultiva en todo el país excluyendo los páramos y sub páramos o zonas de bosques andinos degradados, encima de 3,000 m de altitud, con siembras concentradas en las provincias de Loja, Azuay, Pichincha y, en menor medida, en aquellas de Bolívar, Chimborazo, Tungurahua e Imbabura en la región interandina. Este cultivo es presente en las provincias costaneras de Manabí, seguida por Esmeraldas, Guayas en la región litoral y en la provincia de Pastaza en la región amazónica. El maíz habría cruzado el istmo de Panamá hace 5,000 años a.c., entrando al territorio colombiano, para luego alcanzar la región litoral o costa ecuatoriana.

1.6.11. Naranja fruta fresca (Citrus sinensis)

Los cítricos son una especie que pertenece a la clase Angioespermae, subclase dicotiledónea, orden rutae, la familia rutaceae y género citrus. Existen 145 especies de cítricos y entre ellos la naranja. La variedad de naranja se diferencia según el uso que se le vaya a dar; es decir, si es para consumo fresco o para congelar. Es una de las frutas más populares y saludables del mundo. Tiene un alto contenido de vitamina C. Su sabor, especialmente de algunas variedades es realmente soberbio por su acidez y dulzura.

Aproximadamente un 90 por ciento de su contenido es agua con un cinco por ciento de azúcares. Se considera uno de los cítricos más saludables, pues es rica en nutriente, pero baja en calorías; también, contiene muchos antioxidantes, fibra, vitaminas y minerales. Con todo esto algunos de sus beneficios son al sistema inmune, regula el apetito, mejora la salud cardiaca y ayuda a la pérdida de peso.

La superficie de cultivo de la naranja en Ecuador es de 55.953 hectáreas. En la Provincia Bolívar es de 10.639 hectáreas y en cantón Caluma se cultivan 2.650 hectáreas, que representa el 4, 73% de la producción nacional. Actualmente, el expendio de naranja en región de Caluma se realiza solamente como fruta, sin darle un valor agregado.

Además, la producción y cosecha es estacional; es decir, solo se realiza durante meses de julio, agosto y septiembre del año, durante el resto del año no se obtiene cosecha de dicha fruta. Las variedades de naranja que se cultivan en el cantón son: Valencia tardía, Valencia común, Valencia delta, Thompson, Washington, Naranja lima, Naranja agria y Naranja pomelo. La más consumida y la que se usa en este proyecto es la variedad Valencia común por su aceptabilidad entre los consumidores y su alto contenido en azucares que lo hace ideal para este proceso.

1.6.12. Palma africana fruta fresco (Elaeis guineensis)

La palma es originaria de África occidental, se obtiene aceite desde hace 5,000 años, especialmente en Guinea Occidental de donde pasó a América. Fue introducida después de viajes de Colón y en épocas más recientes fue introducida a Asia desde América. El cultivo en Malasia es importante económicamente, provee la mayor cantidad de aceite de palma y sus derivados a nivel mundial. En América, los mayores productores son Colombia y Ecuador.

El aceite de palma es un aceite de origen vegetal que se obtiene del mesocarpio de fruta de palma. Es el segundo tipo de aceite con mayor volumen de producción, siendo el primero el aceite de soja. Ecuador ocupa el segundo lugar en Latinoamérica en la producción de aceite crudo de palma y es el séptimo productor a nivel mundial, aún con rendimientos más bajos comparados con Colombia y Costa Rica.

A pesar de que los productores de más de 1,000 hectáreas tienen el liderazgo en la industria de la palma, el 87% produce menos de 50 hectáreas. Además, la tasa de deforestación del Ecuador ocupa el noveno puesto en escala de FAO como una de

las más altas del mundo y la más alta de América del Sur; las fincas de palma africana han sido criticadas por estar involucradas en la deforestación y por promover la precarización del trabajo. Sin embargo, sectores del gobierno ven las compañías de palma de aceite como una fuente de empleo y desarrollo para una región pobre.

El trabajo de campo realizado demuestra que hay una diferencia de percepción de pequeños agricultores. Los agricultores de Quinindé-La Concordia se mostraron satisfechos con los ingresos que obtienen y con el alza de los precios de la tierra plantada con palma. Sin embargo, los agricultores de San Lorenzo, no están felices, pues la encuesta demuestra que una enfermedad devastó árboles y, como resultado, sus precios de la tierra en San Lorenzo han caído.

1.6.13. Papa (Solanum tuberosum L)

La papa o patata es un tubérculo comestible que se extrae de la planta herbácea americana ***Solanum tuberosum*** L, de origen andino. Es una planta perteneciente a la familia solanáceas originaria de Suramérica y cultivada por todo el mundo por sus tubérculos comestibles. Fue domesticada en el altiplano andino por sus habitantes entre el 8,500 y el 5,000 a. n. e., y, después, fue llevada a Europa por conquistadores españoles como una curiosidad botánica más que como una planta alimenticia. Su consumo fue creciendo y su cultivo se expandió a todo el mundo hasta convertirse hoy día en uno de los principales alimentos para el ser humano.

La mayor diversidad genética de papa (***Solanum tuberosum*** L) cultivada y silvestre se encuentra en tierras altas de Andes de América del Sur. La primera crónica conocida que menciona la papa fue escrita por Pedro Cieza de León en 1538. Cieza encontró tubérculos que los indígenas llamaban “papas”, primero en la parte alta del Valle del Cuzco, Perú y, posteriormente, en Quito, Ecuador. El centro de domesticación del cultivo se encuentra en los alrededores del Lago Titicaca, frontera actual entre Perú y Bolivia. Existe evidencia arqueológica que prueba que varias culturas antiguas cultivaron la papa; por ejemplo: Inca, Tiahuanaco, Nazca y

Mochica. Actualmente, es el tubérculo andino que los ecuatorianos usan para preparar sus platos más tradicionales como el locro de papas, sopas o caldos, ornado, etcétera. Ayuda al mantenimiento del sistema nervioso, posee alto nivel de complejo B, vitamina C y su índice de calorías es alto; aunque, no más que el arroz, el fideo o el pan.

Según cifras de la Subsecretaría de Agricultura de Ecuador, se producen 421,000 toneladas de papas al año en este país, expertos aseguran que a causa de lluvias y suelo fértil Ecuador es un "paraíso" para cultivar papa y, según Centro Internacional de la Papa, cada ecuatoriano consume 24 Kg/año. Aunque, la mayor producción de papa se da en la sierra por su clima se conoce que el mayor consumo de la papa chola se da en la costa, específicamente en la ciudad de Guayaquil.

Las papas chola y chaucha son las más comerciales, pero su número de variedades en Ecuador es de 570. Muchas de ellas son nativas, siendo su consumo no alto; por ejemplo: uvilla, leona negra, calvache, conejo negro, yema de huevo, caucha colorada, carrizo, coneja blanca y santa rosa. Todas ellas son diferentes en sus formas y colores, pero se asemejan en cantidad de proteínas, fibra y minerales que poseen.

1.6.14. Plátano (Musa paradisiaca)

Pertenece a la familia de las musáceas, que comprende una cincuentena de especies de megaforbas de confusa taxonomía, así como decenas de híbridos. Son plantas muy antiguas, se plantea que es oriunda de la India, donde se encuentra el mayor número de clones. América se conoce como el segundo centro de origen. Este cultivo ocupa a nivel mundial el segundo lugar en consumo fresco, después de cítricos.

Sin embargo, los productores de plátano barraganete de exportación de El Carmen (Manabí) y Santo Domingo de los Tsáchilas se sienten afectados por la caída del precio del producto y enfermedades como la sigatoka, que ataca a las plantaciones. Según la Federación Nacional de Productores de Plátano del Ecuador (Fenaprope),

el precio oficial es de USD 7.30, pero exportadores e intermediarios compran al agricultor a un precio más bajo. "En invierno, la cosecha rinde más que en verano.

Por eso el productor vende a menos precio para que sobreproducción no se dañe". En el país se cosecha para exportar a EE.UU., mercado colombiano y otra parte se queda para el consumo nacional. De las 433 000 toneladas que se produjeron en el 2016, solo el 10% fue para consumo nacional, según el Banco Central. De ese rubro, solo el 2% se procesa.

1.6.15. Tomaté de árbol fruta fresca (Cyphornandra betacea)

Es originaria de Bolivia, Argentina, Perú, Ecuador y Colombia. Se desarrolla en temperaturas que oscilan entre los 14 y 20 grados y entre los 600 y 3 300 metros sobre el nivel del mar. Los frutos se pueden usar en jugos, dulces y postres. Previamente calentadas, el fruto u hojas se aplican en forma tópica contra inflamación de amígdalas o anginas; en caso de gripe, se consume el fruto fresco en ayunas, pues se sabe que el fruto posee alto contenido de ácido ascórbico.

Es considerado uno de los cultivos frutícolas más rentables del Ecuador. En el país se cultivan varios ecotipos, que no son consideradas variedades por expertos; por ejemplo: anaranjado puntón, anaranjado redondo, tomate nacional, amarillo nacional y partenocárpico. El Instituto Nacional Autónomo de Investigaciones Agropecuarias (INIAP) también considera como ecotipos al anaranjado gigante, morado neocelandés y morado gigante. Tungurahua, Pichincha, Imbabura, Cotopaxi, Chimborazo, Azuay y Loja son las zonas del país donde más se acopló el tomate por su clima frío-templado.

En el país se cultivan unas 9 000 hectáreas, principalmente en Tungurahua, Pichincha, Imbabura, Cotopaxi, Chimborazo, Azuay y Loja. 10% de la producción de Tungurahua se envía a Colombia de manera informal.

1.6.16. Trigo (Triticum spp)

Término que designa al conjunto de cereales, cultivados como silvestres, que pertenecen al género Triticum. Son plantas anuales de la familia de gramíneas, ampliamente cultivadas en todo el mundo. La palabra trigo designa tanto a planta como a semillas comestibles. Se piensa que se ha cultivado desde hace más de 9,000 años. Algunos autores piensan que surgió en el valle del Río Nilo.

El trigo entra a América cuando inmigrantes rusos lo trajeron a Kansas en 1873, la variedad llamada Pavo Rojo, que crece mejor que cualquier otra. Chimborazo es una de las provincias con mayor cultivo de trigo en Ecuador, con 1,687 hectáreas, seguida de Bolívar, 1,015 y Pichincha, 685. Según Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC) del Instituto Nacional de Estadística y Censos (INEC), en 2016 Ecuador sembró 4,617 hectáreas de trigo, con 4,422 Ha cosechadas y una producción de 6,746 toneladas. Actualmente, 90 %, 500,000 toneladas, del trigo que consume la industria nacional es importado de Estados Unidos, Canadá y otros países, mientras que 10% restante es producción local.

1.6.17. Yuca raíz fresca (Manihot esculenta Crants)

Yuca, mandioca, guacamota o casava es un cultivo perenne con alta producción de raíces reservantes, como fuente de carbohidratos y follajes para la elaboración de harinas con alto porcentaje de proteínas. Las características de este cultivo permiten su total utilización, el tallo para su propagación vegetativa, sus hojas para producir harinas y las raíces reservantes para el consumo en la alimentación en que se emplea en diferentes platos; también, se usa en elaboración de casabe, agroindustria y exportación.

La yuca o mandioca es una especie de origen americano, que se ha extendido en una amplia área de los trópicos americanos desde Venezuela y Colombia hasta el Noroeste de Brasil, con predominio de los tipos de yuca dulce en el norte y en la zona de Brasil los amargos. La evidencia más antigua del cultivo de la mandioca proviene

de los datos arqueológicos de que se cultivó en el Perú hace 4,000 años y fue uno de los primeros cultivos domesticados en América.

La yuca en Ecuador es un cultivo tradicional que se produce en la costa occidental, amazonia oriental, Loja y Santo Domingo de los Colorados. En Manabí, el mayor porcentaje de productores está constituido por pequeños agricultores de escasos recursos, que la siembran como cultivo de subsistencia sin utilizar tecnologías mejoradas y preferencia intercalada con maíz. Entre las principales limitaciones a producción identificadas están la precipitación irregular, falta de estudios de mercadeo e inadecuada transferencia de tecnología.

El Instituto Nacional de Investigaciones Agropecuarias modifico la tecnología recomendada por CIAT a condiciones ecuatorianas; se han introducido 9 genotipos, actualmente adaptados a condiciones de invernadero; está pendiente la transferencia al campo. Se han realizado pruebas regionales en Quevedo, Santo Domingo, Napo y Loja con variedades e híbridos introducidos y material local.

La estructura para intercambio internacional de germoplasma es aún deficiente. Entre los planes futuros están la colección de material local, evaluación de sus características, mejoramiento de principales sistemas de producción y utilización de yuca en la Provincia de Manabí (CIAT).

1.7. Ejercicios Propuestos

1.7.1. Numéricos

1. Se dese estudiar la relación entre desgaste de hierro dulce y viscosidad de aceite:

Tabla 1. Relación entre desgaste de hierro dulce y viscosidad de aceite

Variable	Datos numéricos									Total
X	1.60	9.40	15.50	20.00	22.00	35.50	43.00	40.50	33.00	220.50

Variable	Datos numéricos									Total
Y	240.00	181.00	193.00	155.00	172.00	110.00	150.00	75.00	94.00	1370.00

Con el propósito de facilitar su estimación, se dan algunos datos generales:

$$S_{xx} = 1{,}651.42, S_{xy} = -5{,}109.60, S_{yy} = 21{,}735.56$$

a) ¿Qué haría para hallar la relación entre dos variables? Enumere por pasos todo su planteamiento y escriba sus argumentos.
b) Estime recta de regresión.
c) ¿Cómo interpreta, en función del objetivo, la recta?
d) Estime coeficiente R^2 y escriba su interpretación.
e) Estime razón F, compárelo con el fractil de orden 0.95 de su correspondiente ley y ¿Qué decisión tomará respecto a significación de regresión?
f) Calcule intervalos de confianza de nivel 95 y 99% para parámetros de regresión.
g) Para $x = 50$, estime intervalos de confianza de nivel 95 y 99% para media y predicción.
h) Realice el gráfico de residuos, ¿considera algún problema en el?, en caso afirmativo escriba ¿qué debe hacer y qué espera como resultado?

2. Un resumen de calificaciones de un grupo de 9 estudiantes en un curso (X) y examen final (Y) es:

Tabla 2. *Resumen de calificaciones de un grupo de 9 estudiantes*

Variable	Datos calificaciones								
Calificaciones (X)	77.00	50.00	71.00	72.00	81.00	94.00	96.00	99.00	67.00
Examen (Y)	82.00	66.00	78.00	34.00	47.00	85.00	99.00	99.00	68.00

Responda a los incisos siguientes:

a) Halle recta de regresión lineal.
b) Estime ANOVA y coeficiente de determinación, ¿qué puede argumentar respecto a bondad de ajuste?
c) Estime intervalos de confianza de nivel 95 y 99% para coeficientes de recta de regresión tal que prueba su nulidad?
d) Calcule ecuaciones de bandas o intervalos de confianza para medias y predicciones.
e) Estime intervalo de confianza, con nivel de confiabilidad estadística 95 y 99 %, para calificación final media de estudiante que obtuvo 85.00 en trabajo de curso?
f) Realice análisis completo de residuos.

3. Con objeto de estimar horas/mes usadas en servicio siderúrgico (Y) en 15 hospitales navales (X), se estimó el modelo con resultados:

$$\frac{1}{y} = \left[\beta_0 + \beta_1 \left(\frac{1}{x}\right)\right]$$

Tabla 3. *Modelo de estimación de horas/mes usadas en servicio siderúrgico* (Y) *en 15 hospitales navales* (X)

Variable	Coeficiente	Razón t
Intercepto (β_0)	-0.00005	-2.715
Variable exógena $\left(\frac{1}{x}\right)$	0.17433	20.253

Resuelva los siguientes incisos:

a) Estime intervalos de confianza con nivel de confiabilidad estadística 95 y 99 % para parámetros.
b) Estime razón F.

c) Estime coeficiente R^2.

d) Argumente si existe o no información suficiente para aceptar o no rechazar el modelo.

4. En el Banco del Pacífico se trata de explicar créditos bancarios (Y) en función de depósitos (X). Se obtienen datos de 25 sucursales con resultados:

$$\bar{x} = 290.07, \bar{y} = 263.19, \sum_{i=1}^{n=25} x_i^2 = 6'964{,}622.95, \sum_{i=1}^{n=25} y_i^2 = 4'832{,}155.27, \sum_{i=1}^{n=25} x_i y_i = 5'472{,}991.60$$

Resuelva los incisos:

a) Escriba modelo de regresión lineal simple con sus hipótesis completas.

b) Calcule estimadores de mínimos cuadrados de parámetros.

c) Anote ecuación de recta de regresión lineal simple obtenida. Si datos están en miles de USD, ¡cómo interpreta sus coeficientes?

d) Pruebe $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ con nivel de confiabilidad estadística 5 y 1 %.

e) Estime intervalo de confianza con nivel de confiabilidad estadística 95 y 99 % para β_2.

f) Calcule tabla ANOVA.

g) ¿Cuál es prueba asociada a razón F, qué relación existe entre F y t-Student?

h) Si probabilidad crítica de razón F es $\hat{\alpha} = 1 * 10^{-10}$, ¿cuál debe ser su decisión correcta?

i) Estime coeficiente de determinación e interprételo.

j) Interprete el siguiente gráfico de residuos:

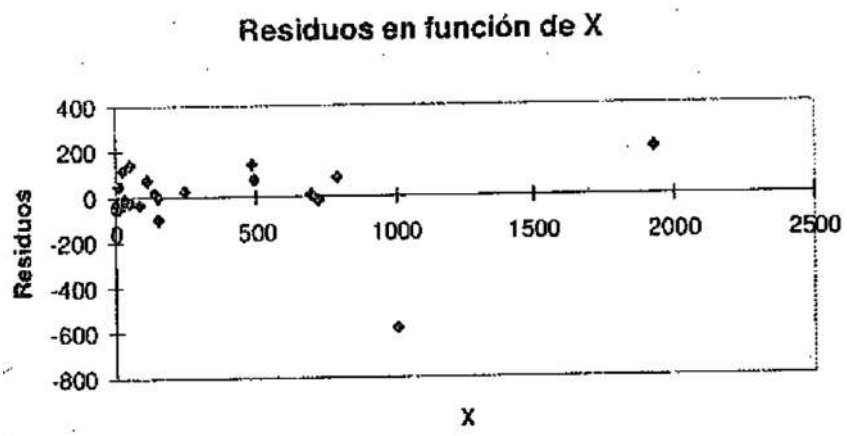

Figura 2. Gráfico de residuos

k) En caso de dos puntos alejados se tiene $x_i = 1{,}921.80, x_j = 1{,}000.00$. Estime valor h_{ii}/punto. Argumente si se pueden considerar como puntos palanca.

l) Estime predicción para en dos puntos atípicos tal que si $y_i = 1{,}672.00, y_i = 200.00$ escriba su argumento respecto cuál puede ser influyente o no.

m) En caso de eliminar el segundo punto atípico, sus resultados son:

$$\hat{y} = 48.89 + 83\,x, R^2 = 0.97$$

$$(t) = (3.21)(0.83)$$

Escriba su argumento respecto a la nueva regresión y punto atípico eliminado.

1.7.2. Teóricos

5. Demuestre que:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})\,x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}$$

n) Recta de regresión para por punto $(\bar{x}, \bar{y})$.

o) Una regresión lineal simple sin término constate:

   a. $\sum_i \hat{u}_i = 0$ ssi $\hat{\beta}_2 = \left(\frac{\bar{x}}{\bar{y}}\right)$.

   b. Estimador:

$$\hat{\beta}_2 = \left(\frac{\sum x_i y_i}{\sum x_i^2}\right)$$

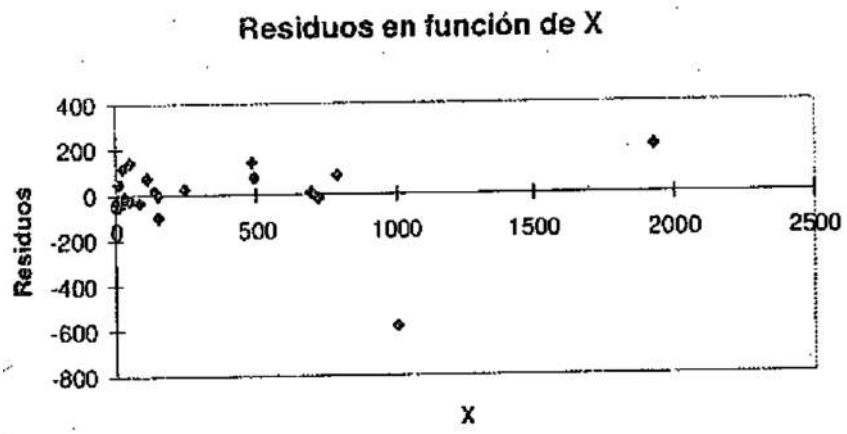

Figura 3. *Regresión lineal simple sin término constate*

Es lineal insesgado y su varianza es $\left(\frac{\sigma^2}{\sum x_i^2}\right)$

c. $\left(\frac{SRC_{(\text{Suma de Residuos al Cuadrado})}}{n-1}\right)$ es estimador insesgado de σ^2.

d. Deduzca $\left(\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\left(\frac{s^2}{\sum_i x_i^2}\right)}}\right)$

p) Demuestre si sesgadas las variables (x_i, y_i) tal que:

a. $\hat{\beta}_1 = 0$ aunque estimación de pendiente $\hat{\beta}_2$ no cambia.

b. Sumas de cuadrados, al igual que relación $STC_{(\text{Suma Total de Cuadrados})} = SEC_{(\text{Suma de Estimaciones al Cuadrado})} + SRC_{(\text{Suma de Residuos al Cuadrado})}$ y tabla ANOVA no cambian.

q) Comprobar que $t^2 = F$ tal que $t = \left(\frac{\hat{\beta}_2}{ee\,(\hat{\beta}_2)}\right)$

r) Demuestre que $F = (n-2)\left(\frac{R^2}{1-R^2}\right)$.

s) Compruebe que si $y_i = c_1$ $i = 1,2,3,4,\ldots,n; y_i = c_2$ $i = n+1, n+2, n+3, n+4, \ldots, n+m$ tal que $c_1 \neq c_2$ donde los puntos $(\hat{y}_i, \hat{u}_i)$ forman dos rectas paralelas de pendiente -1.

6. Se desea estudiar la relación entre la intensidad de regadío (z) y la productividad cierto cultivo. Se obtuvieron los siguientes resultados:

Tabla 4. *Relación entre la intensidad de regadío (z) y la productividad cierto cultivo*

$\mathbf{X_i}$	9	10	13	15	18	13
$\mathbf{V_i}$	36	44	48	63	70	45

Ajuste un modelo lineal simple y calcule el coeficiente de correlación lineal entre las v

7. Se realizó un experimento para medir la velocidad del sonido en el aire a diferentes temí. Los resultados obtenidos son:

Tabla 5. *Velocidad del sonido en el aire a diferentes temí*

Temperatura en °C (i)	-13	0	9	20	33	50
Velocidad en m/s (y)	322	335	337	346	352	365

a. Estime la función de regresión lineal de y sobre x. Interprételos;
b. Calcule el coeficiente de correlación lineal entre las variables;
c. Pruebe si /31 = 0;
d. Encuentre los intervalos de estimación y de predicción cuando la temperatura es

8. Para determinar la relación entre el número de vendedores y las ventas anuales (en dólares) que tiene una empresa, se obtuvieron los siguientes datos:

Número de vendedores	12	13	14	15	16
Ventas	20	27	33	41	58

Se postuló que los datos se ajustan a un modelo lineal simple.

a. Determine las estimaciones de los parámetros /30 y P¡ y escriba el modelo de regó.
b. Grafique los puntos y la recta de ajuste;
c. Construya la tabla de análisis de la varianza y pruebe si /3¡ = 0;
d. ¡Se ajustan bien los datos a la recta?

9. En Una investigación de las propiedades de un pegamento de secado rápido se midió el que se demora en cristalizarse en función de la cantidad de pega depositada sobre una si de material cerámico de prueba.

Cantidad (g)	1.0 1.2 1.4 1.6 1.8 2.0
Tiempo (seg)	26.2 27.9 29.4 30.5 31.0 34.3

a) Ajuste los datos a un modelo lineal simple;
b) Realice un análisis de varianza y pruebe la significación del ajuste;
c) Construya los intervalos de confianza para los parámetros de regresión.

10. En una agencia bancaria se registró el número de depósitos realizados y el monto total transacciones, en una hora de trabajo, con los siguientes resultados.

Tabla 6. *Registró el número de depósitos realizados y el monto total transacciones*

Monto total (en cientos de dólares)	10	5	7	19	11	S
Número de depósitos	16	9	3	25	7	13

a. Realice la formulación matricial de problema y ajuste los datos a un modelo lineal. Interprete los coeficientes;
b. Calcule s y obtenga un intervalo de confianza, al 95 %, para los coeficientes de regresión;
c. Evalúe r2 e interprete su valor. Pruebe si p = 0;

d. Realice una predicción para cuando el número de depósitos es 12;
e. Obtenga la tabla ANOVA y realice la prueba F correspondiente.

11. En el mercado inmobiliario se realiza el avalúo de una propiedad para luego efectuar su venta la diferencia constituye la ganancia del vendedor. En la tabla se dan los valores (en miles de dólares) de avalúo y precio de venta de doce propiedades en Quito.

Tabla 7. *Avalúo de una propiedad para luego efectuar su venta*

Avalúo	Venta	Avalúo	Venta
46.5	59.0	67.4	108.0
43.5	56.5	70.1	95.0
52.2	65.2	74.0	84.0
62.5	74.0	57.4	106.0
85.7	109.0	103.8	154.0

a. Grafique las variables en un diagrama de dispersión;
b. Realice la formulación matricial del problema y encuentre los estimadores de los parámetros del modelo;
c. ¿Aporta el valor de avalúo x información para conocer el precio de venta y?;
d. Calcule r^2. ¿Amerita realizar una prueba de hipótesis para probar si p — 0?;
e. Obtenga un intervalo de confianza de 95 % para el precio medio de una propiedad con un avalúo de 72 mil dólares. Interpreté el intervalo;
f. Encuentre un intervalo de confianza al 95% para el precio de venta de una propiedad avaluada en 65 mil dólares. Interprete el intervalo.

12. En una entidad financiera, se desea tener un método que permita realizar pronósticos de las ganancias obtenidas en base a información que pueda estar disponible de manera rápida. El gerente de crédito plantea un modelo que relaciona el número de préstamos realizados en un mes y la ganancia

obtenida en el mismo período. Para tal efecto recoge la siguiente información de los 8 últimos meses:

Tabla 8. *Modelo que relaciona el número de préstamos realizados en un mes y la ganancia obtenida en el mismo período*

No. préstamos	125	131	142	127	140	121	136	133
Ganancia	44	54	77	35	80	41	66	65

Los valores de las ganancias están en cientos de dólares.

a) Encuentre la ecuación de i egresión lineal simple que relaciona el número de préstamos y la ganancia;
b) ¿Es satisfactorio el modelo obtenido y por qué?
c) Realice un análisis de la varianza;
d) Haga una predicción para un mes en el que se otorgaron .123 préstamos y construya un intervalo al 95 % para tal predicción;
d) ¿Cree que podría mejorarse el modelo propuesto? ¿Cómo?

13. La siguiente tabla muestra la captura de anchoas, millones de toneladas métricas y el precio de harina de pescado, dólares por tonelada, para los últimos 10 años

Tabla 9. *Captura de anchoas, millones de toneladas métricas y el precio de harina de pescado*

Año	1	2	3	4	5	6	7	8	9	10
Precio	190	160	134	129	172	239	542	245	454	410
Captura	7.23	8.53	9.82	10.26	8.96.	4.45	1.78	3.3	0.8	0.5

Construya los modelos lineales que relacionen las variables (x — y) e interprete los coeficientes

a) Precio y año;
b) Captura y año;
c) Precio y captura;

Con el modelo que tenga la máxima correlación:

d) Realice la tabla ANOVA é interprétela;
e) Construya los intervalos de confianza para los coeficientes de regresión;
f) Realice la estimación de y cuando x = 5.5.

14. Los siguientes datos corresponden al ritmo cardiaco en reposo (V) y el peso (X, en kg) personas.

Tabla 10. *Ritmo cardiaco en reposo (V) y el peso (X, en kg) personas*

X	Y
90	62
86	45
67	40
89	55
81	64
75	53
$\sum x_i = 488,$	$\sum y_i = 319,$
$\sum x_i^2 = 40092,$ $\sum x_i y_i = 26184$	$\sum y_i^2 = 17399$

a) Grafique los datos y examine si parece que hay relación lineal entre las dos variantes.
b) Calcule los estimadores de los parámetros de regresión;
c) Obtenga la estimación por intervalo de la media cuando $x = 88$, con 95 % de confiabilidad estadística;
d) Calcule los coeficientes de determinación y de correlación entre las dos variables.

15. En un laboratorio de prueba de automóviles se mide la distancia de frenado en relación cc velocidad que lleva el auto, dando los siguientes resultados de 20 observaciones.

$$x = 50, \qquad \sum_{i=1}^{n}(x_i - x)^2 = 1600, \qquad \sum_{i=1}^{n}(x_i - x)\ (y_i - y = 800,$$

$$y = 30, \qquad \sum_{i=1}^{n}(y_i - y)^2 = 832,$$

a) Calcule la ecuación de ajuste para el modelo de regresión lineal simple;
b) Realice una tabla de análisis de la varianza;
c) Determine los intervalos de confianza de la estimación y de la predicción al 95 %, par valor de la distancia de frenado, cuando la velocidad $x_p = 60$.

16. Se realizó un estudio para determinar el efecto que tiene la temperatura (x) sobre la cantidad de gas residual generando (y) en un proceso químico. Se analizaron 12 unidades muestrales y se midieron las siguientes cantidades.

$$n = 12, \qquad \bar{x} = 7.5, \qquad \hat{y} = 55.1,$$

$$\sum_{i=1}^{12} x_i^2 = 4250, \qquad \sum_{i=1}^{12} y_i^2 = 279360 \qquad \sum_{i=1}^{12} x_i y_i = -23707.5$$

a) Encuentre la ecuación de regresión lineal que explica la cantidad de gas por la temperatura del proceso;
b) Calcule el coeficiente de determinación;

c) Realice un análisis de varianza. Interprételo.

17. Una teoría financiera sostiene que hay una relación directa entre el riesgo de una inversión y el rendimiento que promete. El riesgo de una acción se mide por su valor llamado β. En tabla se muestra los rendimientos y valores de 12 acciones:

Tabla 11. *Rendimientos y valores de 12 acciones*

Rendimiento	5.4	8.9	2.3	1.5	3.7	8.2	5.3	0.5	1.3	5.9	6.8	7.2
Valor Beta	1.5	1.9	1.0	0.5	1.5	1.8	1.3	-0.5	0.5	1.8	1.9	1.9

Al ajustar un modelo de regresión a estos datos se obtuvo:

Tabla 12. *Ajuste de un modelo de regresión a estos datos*

Predictor	Coef.	Error Est.	$t_{Calculada}$
Intercepto	0.44	0.122	3.62
Pendiente	0.19	0.022	8.56

Tabla 13. *ANOVA del ajuste de un modelo de regresión a estos datos*

Fuente de variación	g. l.	SC	MC	F
Regresiòn	–	–	–	–
Residual	–	0.439	–	
Total	11	3.649		

a) Pruebe si los coeficientes del modelo son significativos. escriba las pruebas utilizadas con sus valores y las regiones de rechazo;
b) Complete la tabla ANOVA e interprétala;
c) Calcule el coeficiente de determinación e interprétalo.

18. Suponga que se ajustado una línea recta a un conjunto de 9 parejas de observaciones, dando:

$$\hat{y} = -5 + 2x$$

Además, se obtuvieron las siguientes desviaciones: $(x_i - \bar{x})$: $-4, -3, -2, -1, 0, 1, 2, 3, 4$ y análisis de la varianza:

Tabla 14. *ANOVA del ajuste de un modelo de regresión a estos datos*

Fuente de variación	g.l	SC	MC	F
Regresión	–	–	–	–
Residual	–	–	–	
Total	–	380		

a) Complete la tabla ANOVA;
b) Realice una prueba de adecuación de modelo.

Modelos no lineales:

Se presente 7 mediciones de dos variables:

Tabla 15. *Mediciones de dos variables*

x	0.5	1.0	1.5	2.0	2.5	3.0	3.5
y	0.6	2.7	12.2	54.6	244.7	1096.6	4914.8

Encuentre la ecuación de regresión que ajusta los datos, según un modelo exponencial y c el coeficiente de determinación.

19. En la siguiente tabla se encuentra el número de años para el vencimiento y el rendimiento de unos bonos.

Tabla 16. *Número de años para el vencimiento y el rendimiento de unos bonos*

Años para el vencimiento	1	2	5	10	15	18	23	
Rendimiento	0.067	0.072	0.076	0.079	0.081	0.077	0.082	0.

a) Ajuste un modelo multiplicativo a los datos;

b) Ajuste un modelo $\hat{y} = b_0 + \frac{b_1}{t}$;

c) ¿Cuál de los dos modelos cree es mejor y por qué?

20. El gerente de una empresa desea relacionar la evolución de las ventas y el gasto publicitario según los datos que aparecen en el cuadrado:

Tabla 17. *Relación de la evolución de las ventas y el gasto publicitario*

Ventas (y)	10	15	18	23	25
Gastos (y)	19	22	41	72	98

a) Realice un ajuste de tipo multiplicativo y encuentre de la calidad de ajuste;

b) Ajusta los datos mediante una función lineal simple y compare con el ajuste anterior;

c) ¿Cuál de los dos modelos es mejor? Explique.

21. Considérese los datos que se presenta a continuación

Tabla 18. *Datos para ejercicio*

x	9	14	17	11	8	10	5	7
y	0.34	0.26	0.18	0.30	0.40	0. 27	0.27	0.42

a) Grafique los datos;

b) Supongo que las variables x i y se vinculan mediante la relación $y = \frac{1}{\beta_0 + \beta_{1x} + \epsilon}$. Encuentre los coeficientes la regresión.

c) ¿El modelo es adecuado para los datos? ¿Por qué?

22. Enseguida se presenta la evaluación anual del salario mínimo vital de un país con alto índice de inflación.

Tabla 19. *Evaluación anual del salario mínimo vital de un país con alto índice de inflación*

Año	1	2	3	4	5	6	7	8	9	10	11	12
S.M. V	66	95	120	120	145	190	220	320	320	400	600	600

a) Grafique los datos;
b) Ajuste los datos mediante un modelo exponencial;
c) ¿Qué se puede decir de la calidad de ajuste de los datos a la curva de regresión?;
d) Realice una estimación de valor del S.M. V. en julio de 1994;
e) Si el S.M.V. en el año 13 ascendió a 900, ¿es una buena la predicción realizada? Explique.

23. En astronomía se denomina año sideral al número de años terrestres que un planeta se demora en completar una revolución alrededor del sol y depende de la distancia entre los dos astros. En la tabla se muestra la distancia promedio y el año sideral para los planetas del sistema solar. emplear los datos para determinar un modelo de regresión que relacione las dos variables, tomando como la dependiente al año sideral. (para realizar la trasformación adecuada refiérase a la tercera ley de Kepler)

Tabla 20. *Distancia promedio y año sideral para los planetas del sistema solar*

Planeta	**Distancia al Sol (r: millones de km)**	**Duración del año sideral (T : años terrestres)**
Mercurio	58	0.241
Venus	108	0.615
Tierra	150	1.000
Marte	228	1.880
Júpiter	778	11. 862
Saturno	1428	29. 458
Urano	2871	84. 018
Neptuno	4500	164. 780
Plutón	5913	248. 400

24. Los siguientes datos corresponden al precio de venta (en cientos de dólares) de un modelo de automóvil, según los años de uso.

Tabla 21. *Precio de venta de un modelo de automóvil*

Años de uso	0	1	2	3	4	5	6
Precio	10.2	8.3	6.9	5.5	4.0	3.5	3.3

Ajuste a los datos un modelo recíproco y calcule el coeficiente de determinación.

25. Una empresa de telefonía celular ha registrado la siguiente evaluación en el número de abonados a su servicio en sus primeros 8 años de operación:

Tabla 22. *Evaluación en el número de abonados a su servicio en sus primeros 8 años de operación*

Año	Abonados
1	32000
2	37500
3	41000
4	58000
5	107000
6	138000
7	175000
8	321500

a) Grafique los datos;
b) Se planteó que el modelo lineal simple. Realice la estimación de los coeficientes;
c) ¿Qué se puede decir de la calidad de ajuste de los datos?,
d) Efectué las pruebas de hipótesis sobre los coeficientes del modelo. sí es posible excluir algunos de ellos, ¿Cómo quedaría el modelo final?;
e) Los datos sugieren una relación del tipo exponencial. Realice un ajuste de los datos al modelo propuesto;

Encuentre el coeficiente de determinación y compárelo con el modelo anterior. ¿Cuál es mejor? ¿Por qué?

CAPÍTULO II

CAPÍTULO II

REGRESIÓN LINEAL MÚLTIPLE

2.1 Modelo

Su finalidad es igual a regresión lineal simple; aunque, con la diferencia que interviene más de un regresor en el modelo tal que en lugar de rectas de regresión se tendrán hiper planos de regresión. Un plano en un espacio de tres dimensiones para el caso de dos regresores. En este tema se abordan temas estimación, inferencia, adecuación de modelo y predicción.

No siempre un regresor es suficiente para explicar todas las variaciones de una variable dependiente (Y); por ejemplo: el gerente de una empresa desea incrementar las ventas tal que decide realizar gastos en publicidad y medir la variación en sus ventas. Inicialmente, decide puntuar la publicidad en televisión, pero posteriormente decide también ponerla en la radio y los periódicos.

En la primera etapa la variable de respuesta, que es el incremento en las ventas depende de una sola variable predictora (gastos en televisión) y para realizar un análisis es suficiente empleando un modelo de regresión lineal simple. Más en la segunda etapa, si variable de respuesta depende de un sólo análisis no es suficiente la regresión lineal simple.

En general, aunque hay muchos problemas prácticos que atañen a variables predictoras únicas es mucho más frecuente que la variable repuesta dependa de un conjunto de variables predictoras o transformaciones de las mismas.

Suponga que Y es una variable que depende funcionalmente de ciertas variables $X_2, X_3, X_4, X_5, \dots, X_k$ tal que la numeración comienza en dos pues el número uno se reserva para el término constante o intercepto. La relación exacta entre Y y $\{X_j\}$ se desconoce, pero supone pertenece a una familia de funciones F llamada modelo.

El más usado y sirve como punto de partida para hallar otros es el lineal tal que supone que F es una familia de funciones lineales afines:

$$F = \{Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots + \beta_k X_k | \beta_1, \beta_2, \beta_3, \beta_4, \dots, \beta_k \in \mathbb{R}\}$$

Se sabe que variable Y toma nombres como dependiente, endógena, respuesta o explicada mientras que variables $\{X_j\}$ toman nombres independientes, exógena, regresora, predictor o explicativa. No obstante, para determinar aproximadamente la función que liga a variable dependiente Y con variables independientes $\{X_j\}$ se realizan n observaciones de $k-$ uplas:

$$(y_i, x_{i2}, x_{i3}, x_{i4}, \dots, x_{ik})\ i = 1,2,3,4, \dots, n$$

Se conoce que y_j es valor que toma variable Y cuando $X_2 = x_{i2}, x_{i3}, x_{i4}, \dots, X_k = x_{ik}$. El modelo de regresión lineal múltiple supone que y_i es una observación de una variable aleatoria Y. Las variables $\{X_j\}$ se estima determinísticas o no aleatorias, pues sus valores pueden ser fijados según conveniencias del experimentador.

En una investigación relacionada con experimento para cada $k-1$ upla $(x_{i2}, x_{i3}, x_{i4}, \dots, x_{ik})$ se hacen varias observaciones o mediciones de variable respuesta o endógena Y. No obstante, en econometría es común trabajar con datos provenientes de una muestra, en este caso todas las variables que intervienen en modelo pueden ser aleatorias, pero variables exógenas participan como si fueran determinísticas tal que sus valores no dependen del azar o aleatoriedad.

El modelo de regresión lineal múltiple, que liga a una variable dependiente Y con k variables independientes, exógenas, explicativas mediante la ecuación:

$$y_i = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \cdots + \beta_k x_{ik} + u_i\ \ i = 1,2,3,4, \dots, n$$

se llama modelo de regresión lineal múltiple con k variables independientes, exógenas, regresoras, predictoras o explicativas, $\beta_k\ \ k = 1,2,3,4, \dots, j$ coeficientes de regresión y, también, parámetros desconocidos, no aleatorios, estimados con base en datos observados. Paralelamente, el caso de una sola variable se considera que el

error o perturbación $u_i \equiv e_i$ tiene esperanza 0, varianza σ^2, errores e_i por observación, no correlacionados y es una variable no observable.

Observación. El término lineal refiere a que la relación es como tal en parámetros, pero no en variables. Las técnicas de regresión lineal múltiple pueden usarse si la relación es lineal en parámetros, pero no se olvide que se busca una función, no necesariamente lineal, que relacione variables Y con $\{X_j\}$:

$$Y_{(\text{dependiente,endógena,respuesta o explicada})} = f(X_2, X_3, X_4, X_5, \dots, X_k)$$

Sin embargo, esta relación es ambigua o incierta, pues se ve acompañada por errores y depende de algunos parámetros desconocidos. Además, las hipótesis o supuestos en que se sustenta la regresión lineal múltiple son iguales que regresión lineal simple.

Con el objeto que estimadores de mínimos cuadrados sean únicos se supones que los siguientes vectores son linealmente independientes:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \\ \vdots \\ x_{n2} \end{pmatrix}, \dots, \begin{pmatrix} x_{1k} \\ x_{2k} \\ x_{3k} \\ x_{4k} \\ \vdots \\ x_{nk} \end{pmatrix}$$

Repaso a Álgebra Lineal:

1. Vectores $(u_1, u_2, u_3, u_4, \dots, u_k)$ de $\mathbb{R}^n$ son linealmente independientes ssi:

$$a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4 + \cdots + a_ku_k = 0$$

   Implica que todos los coeficientes $a_j = 0$ o nulos. Si no, son linealmente independientes. Un conjunto infinito B de vectores es linealmente independiente ssi todo conjunto B de vectores es linealmente independiente ssi todo sub conjunto finito de B es linealmente independiente.

2. Si vectores $\{u_j\}$ son linealmente dependientes existe, en igualdad anterior, un coeficiente $a_j \neq 0$ o no nulo tal que el vector correspondiente a dicho coeficiente es combinación lineal del resto de vectores. Por ejemplo: $a_j \neq 0 \Rightarrow u_k = \left(-\frac{1}{a_k}\right)(a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4 + \cdots + a_{k-1}u_{k-1})$.

3. Bases y dimensión. Sea E un espacio vectorial y B un sub conjunto de E:

   a) Si B genera a E ssi ∀ vector de E es una combinación lineal de vectores de B.

   b) Si B es linealmente independiente se dice que B es una base de E.

   c) Si B es una base de E, el número de elementos de B se llama dimensión de B, si B es infinito se dice que E es dimensión infinita.

4. Número de filas linealmente independientes, en una matriz $A_{(n*m)}$ es igual a número de columnas linealmente independientes.

5. Sea A una matriz $A_{(n*m)}$. El número de filas o columnas linealmente independientes se llama rango de matriz y se denota como $rg(A_{(Rango)})$.

6. De bases y dimensión. Sea E un espacio vectorial y B un sub conjunto de E, es obvio que $rg(A_{(Rango)}) \leq \min\{n, m\}$. Si se logra igualdad $[rg(A_{(Rango)}) \leq \min\{n, m\}]$ se dice que matriz A es rango completo.

7. Matriz cuadrada A es simétrica ssi al cambiar filas por columnas la matriz A no cambia; es decir, ssi $A^t = A$ donde A^t es transpuesta de A.

8. Matriz cuadrada A es diagonal ssi es cuadrada y todos los elementos fuera de la diagonal principal son 0 o nulos:

$$A = \text{diag}_{(\text{Diagonal})}\{a_1 + a_2 + a_3 + a_4 + \cdots + a_n\} = \begin{bmatrix} a_1 & 0 & 0 & 0 & 0 & & 0 \\ 0 & a_2 & 0 & 0 & 0 & & 0 \\ 0 & 0 & a_3 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & a_4 & 0 & & 0 \\ 0 & 0 & 0 & 0 & a_5 & & 0 \\ & & \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & a_n \end{bmatrix}$$

9. Matriz identidad es matriz diagonal con todos $a_i = 0$:

$$I = \text{diag}_{(\text{Diagonal})}\{1,1,1,1, \dots ,1\}$$

Cuando requiere indicar el orden de matriz se estribe I_n en vez de I.

10. Si A matriz cuadrada $(n * n)$ es no singular ssi existe una matriz B cuadrada $(n * n)$:

$$AB = BA = I$$

Matriz B se llama inversa de A y se escribe A^{-1}

11. $(AB)^t = B^t A^t$. Si matrices (A, B) son no singulares y de igual dimensión $(AB)^{-1} = B^{-1}A^{-1}$.

12. Una matriz cuadrada $A_{(n*n)}$ es simétrica semi definida positiva, escrita como $A_{(n*n)} \geq 0$, ssi es simétrica y $\forall x, x \in \mathbb{R}^n, x^t Ax \geq 0$ se dice simétrica definida positiva, escrita $A > 0$, ssi es se mi definida positiva y $x^t Ax = 0 \Rightarrow x = 0$.

13. Sea E, F dos espacios vectoriales sobre un mismo cuerpo de escalares K. Una función $f: E \longrightarrow F$ se dice aplicación lineal ssi $\forall a \in k, \forall x, y \in E$ se tiene que $f(ax + y) = af(x) + f(y)$.

14. $\forall A_{(m*n)}$ define una aplicación lineal $\mathbb{R}^n$ en $\mathbb{R}^m$:

$$A: \mathbb{R}^n \rightarrow \mathbb{R}^m: x \rightarrow Ax$$

A su vez toda aplicación lineal de $\mathbb{R}^n$ en $\mathbb{R}^m$ define una matriz $(m * n)$.

15. Si A es aplicación lineal de $\mathbb{R}^n$ en $\mathbb{R}^m$ se llama espacio imagen de A al conjunto:

$$\mathrm{img}_{(\mathrm{Imagen})}(A) = \{w \in \mathbb{R}^m | w = Au, u \in \mathbb{R}^n\}$$

Se llama núcleo de A al conjunto:

$$\mathrm{Ker}(A) = \{u \in \mathbb{R}^n | Au = 0\}$$

Se puede demostrar que $\dim_{(\mathrm{Dimensión})}\left[\mathrm{img}_{(\mathrm{Imagen})}(A)\right] = \mathrm{rg}\left(A_{(\mathrm{Rango})}\right)$ y $\dim_{(\mathrm{Dimensión})}\left[\mathrm{img}_{(\mathrm{Imagen})}(A)\right] + \dim_{(\mathrm{Dimensión})}\left[\mathrm{Ker}_{(\mathrm{Kernel\ o\ núcleo})}(A)\right] = n$

16. Si $A_{(n*n)}$ es matriz cuadrada $(n * n)$ simétrica definida positiva, se define el producto escalar de dos vectores u, v de $\mathbb{R}^n$ respecto de A por $\langle u, u\rangle_{(A)} = (u^t)(Au)$ se define norma de v respecto de A por $\|u\|_{(A)} = \sqrt{(u^t)(Au)}$. Si A es matriz identidad no hace falta sub índices tal que se dice que son producto escalar y norma usuales.

17. Si A es matriz cuadrada $(n * n)$ se llama traza de A a suma de elementos de su diagonal principal tal que $\mathrm{tr}_{(\mathrm{Traza\ de\ A})}(A) = \sum_{i=1}^{n}(a_{ii})$.

Sea A matriz $(n * m)$, B matriz $(m * n)$ tal que $\mathrm{tr}_{(\mathrm{Traza\ de\ AB})}(AB) = \mathrm{tr}_{(\mathrm{Traza\ de\ BA})}(BA)$.

2.2 Nomenclatura matricial.

$$Y_{(\mathrm{dependiente,endógena,respuesta\ o\ explicada})} = \underset{(n*1)}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{bmatrix}}$$

$$X_{(\text{independientes,exógena,regresora,predictor o explicativa})} = \begin{bmatrix} 1 & x_{12} & x_{13} & x_{14} & x_{15} & & x_{1k} \\ 1 & x_{22} & x_{23} & x_{24} & x_{25} & & x_{2k} \\ 1 & x_{32} & x_{33} & x_{34} & x_{35} & \cdots & x_{3k} \\ 1 & x_{42} & x_{43} & x_{44} & x_{45} & & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ & & \vdots & & & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & x_{n4} & x_{n5} & \cdots & x_{nk} \end{bmatrix}_{(n * k)}$$

$$\beta_{(\text{coeficientes de regresión,parámetros desconocidos no aleatorios})} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \vdots \\ \beta_k \end{bmatrix}_{(k * 1)}$$

$$u_i \equiv e_{i_{(\text{error aleatorio o perturbación estocástica})}} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ \vdots \\ u_k \end{bmatrix}_{(n * 1)}$$

Con base en esto, el modelo $y_i = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \cdots + \beta_k x_{ik} + u_i \;\; i = 1,2,3,4, \ldots, n$ se escribe:

$$Y_{(\text{endógena})} = X_{(\text{exógena})}\beta_{(\text{coeficientes de regresión})} + e_{i_{(\text{error estocástico})}}$$

Las hipótesis respecto a errores se reducen a:

$$E\left[u_{i_{(\text{error estocástico})}}\right] = 0, \text{Var}\left[u_{i_{(\text{error estocástico})}}\right] = \sigma^2 I_{(\text{Matriz identidad } n*n)}$$

Las hipótesis (N) se reducen a $u_{i_{(\text{error estocástico})}} \longrightarrow N_n(0, \sigma^2 I)$ justificada debido a que la distribución conjunta de n variables aleatorias o estocásticas independientes $\{u_i\}$ con ley $N(0, \sigma^2)$ es normal multivariante de parámetros $(0, \sigma^2 I)$ tal que $0 \in \mathbb{R}$. El rango completo señala que columnas de matriz X son linealmente independientes.

2.3 Estimación de parámetros

Al igual que en el método de regresión lineal simple, para la estimación de los parámetros se aplica el método de mínimos cuadrados. Suponga que disponemos de $n > k$ observaciones, se denota como x_{ij} al valor de la $i-$esima observación de la variable x_i, como se observa en:

y	x_1	x_2	…	x_k
y_1	x_{11}	x_{12}	…	x_{1k}
y_2	x_{21}	x_{22}	…	x_{2k}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
y_n	x_{n1}	x_{n2}	…	x_{nk}

Si $\hat{y}$ es la predicción de y, la ecuación de regresión queda:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k,$$

Donde $b_0, b_1, \ldots \ldots, b_k$ son tales que la suma de los cuadrados de las diferencias entre los valores observados de la variable respuesta y su estimación por la ecuación de regresión sea mínima. Si se escribe la igualdad para cada una de las observaciones:

$$\hat{y}_1 = b_0 + b_1x_{11} + b_2x_{12} + \cdots + b_kx_{1k}$$

$$\hat{y}_1 = b_0 + b_1x_{21} + b_2x_{22} + \cdots + b_kx_{2k}$$

$$\begin{matrix} . & . \\ . & = . \\ . & . \end{matrix}$$

$$\hat{y}_n = b_0 + b_1x_{n1} + b_2x_{n2} + \cdots + b_kx_{nk},$$

O:

$$\hat{y}_i = b_0 + \sum_{j=1}^{k} b_j x_{ij,} \qquad i = 1,2, \dots , n$$

Se minimizará la suma de los cuadrados de los errores:

$$SCE = \sum_{i=]}^{n} (y_i - \hat{y}_1)^2$$

Derivando SCE con respecto a cada a cada b_i, e igualando el resultado acero se obtiene las ecuaciones:

$$\frac{\partial(SCE)}{\partial b_0} = -2 \sum_{i=]}^{n} (y_i - \sum_{j=1}^{k} b_j x_{ij,}) = 0$$

$$\frac{\partial(SCE)}{\partial b_1} = -2 \sum_{i=]}^{n} x_{i]} (y_i - b_0 - \sum_{j=1}^{k} b_j x_{ij,}) = 0$$

. .
. = .
. .

$$\frac{\partial(SCE)}{\partial b_1} = -2 \sum_{i=1}^{n} x_{ik} (y_i - b_0 - \sum_{j=1}^{k} b_j x_{ij,}) = 0$$

Luego de simplificar las igualdades se obtiene las ecuaciones normales de mínimos cuadrados:

$$nb_0 + b_1 \sum_{i=1}^{n} xi_] + b_1 \sum_{i=1}^{n} xi_2 + \cdots + b_k \sum_{i=1}^{n} xi_k = \sum_{i=1}^{n} 1/1$$

$$b_0 \sum_{i=1}^{n} xi_1 + b_1 \sum_{i=1}^{n} x^2{}_1 + b_1 \sum_{i=1}^{n} x_{i1} x_{12} + b_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1}\, 1/1$$

$$b_0 \sum_{i=1}^{n} xi_k + b_1 \sum_{i=1}^{n} x_{ik}x_{i1} + b_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + b_k \sum_{i=1}^{n} x^2{}_{ik} = \sum_{i=1}^{n} x_{ik}\ 1/1$$

Se dispone de un sistema de $k+1$ ecuaciones normales que involucran a los coeficientes conocidos. Su solución permite conocer los estimadores de los parámetros del modelo lineal; aunque, se observa que es muy laboriosa.

Propiedad de b. El estimador de mínimos cuadrados b tiene las siguientes propiedades:

1. Es insesgado para β.

   Puesto que $(E) = 0$ y $(X^tX)^{-1}X^tX = I$ se tiene:

   $$E(b) = [(X^tX)^{-1}X^tY] = [(X^tX)^{-1}X^t(X\beta + \epsilon)]$$

2. La matriz de covarianza d viene dada por:

   $$Cov\ (b) = a^2(X^tX)^{-1}$$

La matriz Cov (b) es asimétrica además el valor a^2suele ser desconocido debiendo ser estimado.

Estimación de σ². Considere ahora la suma de los cuadrados de los errores:

$$SCE = \sum_{i=1}^{n}(yi - \hat{y}i)\ \epsilon^t\epsilon,$$

Que después de sustituir el valor de E se ilega

$$SCE = YtY — btXtY,$$

que tiene n — k grados de libertad. Así, el error cuadrático medio, que es un estimador insesgado de σ^2, se calcula por:

$$MCE = S^2 = \frac{SCE}{n - k - 1}$$

2.4 Coeficientes de regresión estandarizados

Cuando se realiza un modelo de regresión múltiple, generalmente, las unidades de medición no son las mismas para la variable dependiente y para las variables independientes; de manera que los coeficientes de regresión no se pueden comparar directamente. Para superar esta dificultad, se emplean los coeficientes de regresión estandarizados beta.

Las unidades de medición de todas las variables se transforman a unidades de desviación estándar, dividiendo cada variable por su desviación estándar. Por ejemplo, en la ecuación de regresión se tiene:

$$\frac{\hat{y}}{s_y} = \frac{b_o}{s_y} + \left(b_1 \frac{s_{x1}}{s_y}\right) \frac{x_1}{s_{x1}} + \left(b_2 \frac{s_{x2}}{s_y}\right) \frac{x_2}{s_{x2}}$$

Los coeficientes:

$$\beta_i = b_i \left(\frac{s_{x_i}}{s_y}\right)$$

Son los coeficientes de regresión parcial estándar y su interpretación es la siguiente: si hay una variación de una desviación estándar en x_i, habrá una desviación de betas desviaciones estándar en y.

2.5 Estimación puntual

Si $(a^t)(a)\text{Var}(\hat{\theta}) \geq \text{Var}(a^t\hat{\theta})$ se deduce $\hat{\theta}$ es mejor que $\tilde{\theta}$ si y solo si varianza de toda combinación lineal de $\hat{\theta}$ es menor que varianza de respectiva combinación lineal de $\tilde{\theta}$.

Teorema. Según teorema de Gauss-Markov bajo hipótesis H, el $\text{EMC}_{(\text{Estimador de Mínimos Cuadrados})}$ $(\hat{\beta})$ es el mejor estimador lineal insesgado de β.

Demostración. Sean $\mathcal{C} \in \mathbb{R}^k - \{0\}, \psi_{(\text{Letra griega Psi})} = \mathcal{C}^t\beta, \hat{\beta}$ en $\text{EMC}_{(\text{Estimador de Mínimos Cuadrados})}$ de β, suponga:

$$\widehat{\Psi}_{(\text{Letra griega Psi})} = \mathcal{C}^t\hat{\beta} = \mathcal{C}^t[(X^tX)^{-1}(X^tY)] = b^tY \text{ tal que } b = [(X^tX)^{-1}(X\mathcal{C})]$$

Con base en esto, $\widehat{\Psi}_{(\text{Letra griega Psi})}$ es estimador lineal de $\Psi_{(\text{Letra griega Psi})}$; aunque, por insesgado se tiene:

$$E\left(\widehat{\Psi}_{(\text{Letra griega Psi})}\right) = \mathcal{C}^t[(X^tX)^{-1}X^tE(Y)] = \mathcal{C}^t\beta$$

Si $\widehat{\Psi}_{(\text{Letra griega psi})} = a^tY$ es otro estimador lineal insesgado de $\Psi_{(\text{Letra griega Psi})}$ se tiene:

$$E\left(\widehat{\Psi}_{(\text{Letra griega Psi})}\right) = a^tX\beta = \mathcal{C}^t\beta$$

Aunque, por inyectividad de aplicaciones: $\beta \longrightarrow a^tX\beta, \beta \longrightarrow \mathcal{C}^t\beta$ tal que $\mathcal{C}^t = a^tX$. Por definición $b = [(X^tX)^{-1}(X\mathcal{C})]$ y $\mathcal{C}^t = a^tX$:

$$\begin{aligned}(a-b)^tb = a^tb - b^tb &= [(a^tX(X^tX)^{-1}\mathcal{C}) - (\mathcal{C}^tX^tX(X^tX)^{-1}X(X^tX)^{-1}\mathcal{C})] \\ &= [(a^tX(X^tX)^{-1}\mathcal{C}) - (a^tX^tX(X^tX)^{-1}X(X^tX)^{-1}\mathcal{C})] \\ &= [(a^tX(X^tX)^{-1}\mathcal{C}) - (a^tX(X^tX)^{-1}\mathcal{C})] = 0\end{aligned}$$

Si compara varianzas de estimadores es:

$$\begin{aligned}a^ta = (a-b+b)^t(a-b+b) &= (a-b)^t(a-b) + 2(a-b)^tb + b^tb \\ &= (a-b)^t(a-b) + b^tb \geq b^tb \Rightarrow \text{Var}(a^tY) - \text{Var}(b^tY) = \sigma^2(a^ta - b^tb) \\ &\geq 0\end{aligned}$$

Observación. $\hat{\beta}$ es estimador de varianza mínima, entre estimadores lineales insesgados, se dice ser un estimador eficiente. La eficiencia se da sólo entre estimadores lineales insesgados, que significa que pueden existir otros estimadores que no so lineales o no son insesgados, pero su error cuadrado medio $E\left[(\hat{\beta}-\beta)^t(\hat{\beta}-\beta)\right]$ es menor.

Corolario. Bajo hipótesis H, si matriz X es de rango completo y no aleatoria:

a. $\widehat{Y} = X\hat{\beta} = HY(H = X^tX(X^tX)^{-1})$ es estimador lineal insesgado y eficiente de $E[Y] = X\beta$.
b. $\text{Var}\left[\widehat{Y}\right] = \sigma^2H$.

Demostración. Toda combinación lineal de $Y = X\beta$ es combinación lineal de β; aunque, por Teorema de Gauss-Markov son de varianza mínima. También:

$$\mathrm{Var}\left[\widehat{Y}\right] = \sigma^2 XX^t\left[\mathrm{Var}\left(\hat{\beta}\right)\right] = \sigma^2 XX^t[X(X^tX)^{-1}] = \sigma^2 H$$

Teorema. Bajo hipótesis H, si matriz X es de rango completo y no aleatoria:

a. $E\left[\widehat{U}\right] = 0$
b. $\mathrm{Var}\left[\widehat{U}\right] = \sigma^2(I - H)$

Teorema. Bajo hipótesis N, si matriz X es de rango completo k y no aleatoria:

a. $\hat{\beta} \longrightarrow N_k[\beta, \sigma^2(X^tX)^{-1}]$
b. $\widehat{Y} \rightarrow N_n[X\beta, \sigma^2 H]$ tal que $H = XX^t[X(X^tX)^{-1}]$
c. $\widehat{U} \rightarrow N_n[0, \sigma^2(I - H)]$
d. $\hat{\beta}$ y $\widehat{U}$ son independientes
e. $\left[\left(\frac{1}{\sigma^2}\right)(\widehat{U}^t)(\widehat{U})\right] = \left[\left(\frac{1}{\sigma^2}\right)\left(\mathrm{SRC}_{(\text{Suma de Residuos al Cuadrado})}\right)\right] \rightarrow \chi^2_{(n-k)}$

Demostración. Sea $A = X^t[(X^tX)^{-1}], H = XA$ tal que si $\hat{\beta} \longrightarrow N_k[\beta, \sigma^2(X^tX)^{-1}]$ y $\widehat{U} \rightarrow N_n[0, \sigma^2(I - H)]$:

$$\begin{pmatrix}\hat{\beta}\\ \widehat{U}\end{pmatrix} = \begin{pmatrix}A\\ I - H\end{pmatrix} Y \text{ tal que } Y \rightarrow N_n[X\beta, \sigma^2 I]$$

Y X es vector aleatorio de distribución $N_n(\mu, \Sigma)$, donde A es una matriz $m * n \Rightarrow AX \longrightarrow N_m(A\mu, A\Sigma A^t)$:

$$\begin{pmatrix}\hat{\beta}\\ \widehat{U}\end{pmatrix} \longrightarrow N_{n+k}\left[\begin{pmatrix}A\\ I - H\end{pmatrix} X\beta, \sigma^2 \begin{pmatrix}AA^t & A(I - H)\\ (I - H)A^t & (I - H)(I - H)^t\end{pmatrix}\right]$$

Por teorema de Gauss-Markov bajo las hipótesis H. Sea $\psi = a\beta_1 + b\beta_2$ donde $a, b \in \mathbb{R}$. Si $\hat{\beta}_1, \hat{\beta}_2$ son estimadores de MC, entonces $\widehat{\psi} = a\hat{\beta}_1 + b\hat{\beta}_2$ es estimador lineal e insesgado de ψ, pero aún más es de menor varianza entre todos los estimadores lineales e insesgados. Entonces, $\widehat{\psi}$ es mejor estimador lineal insesgado (BLUE) o estimador lineal insesgado eficiente:

$$\begin{pmatrix} A \\ I - H \end{pmatrix} X\beta = \begin{pmatrix} \beta \\ 0 \end{pmatrix}$$

Si se multiplican matrices sería:

$$\begin{pmatrix} AA^t & A(I-H) \\ (I-H)A^t & (I-H)(I-H)^t \end{pmatrix} = \begin{pmatrix} (X^tX)^{-1} & 0 \\ 0 & (I-H) \end{pmatrix}$$

Por propiedad $X \longrightarrow N_n(\mu, \Sigma)$ si X se parte en dos sub vectores de m y $n-m$ componentes, respectivamente, si μ y Σ se hacen particiones similares tal que $X = \begin{pmatrix} X^1 \\ X^2 \end{pmatrix}, \mu = \begin{pmatrix} \mu^1 \\ \mu^2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma_{(21)} & \Sigma_{(22)} \end{pmatrix}$ tal que $X^1 \longrightarrow N_m(\mu^1, \Sigma_{(11)})$ aunque se puede generalizar a cualquier sub vector e incluso las leyes marginales de un vector aletorio de ley normal son normales con sus respectivas medias y varianzas-covarianzas, $\hat{\beta}$ y $\hat{U}$ son normales; además, por propiedad, en que notaciones e hipótesis anteriores, si $Cov(X^1, X^2) = \Sigma_{(12)} = 0$ los vectores X^1 y X^2 son independientes tal que para sub vectores de un vector de ley normal la no correlación equivale a independencia.

Aunque, la última es consecuencia del Teorema de Cochran.

Sean: X es un vector aleatorio de ley $N_n(\mu, \Sigma), \Sigma$ no singular; $E_1, E_2, E_3, E_4, \dots, E_k$ sub espacios vectoriales de $\mathbb{R}^n$ dos a dos Σ^{-1} ortogonales tal que $\mathbb{R}^n = E_1 \oplus E_2 \oplus E_3 \oplus E_4 \oplus \dots \oplus E_k$; X_j la proyección Σ^{-1} ortogonal de X sobre E_j, μ_j la proyección Σ^{-1} ortogonal de μ sobre E_j. Con base en esto se tiene: 1. Vectores X_j son normales con media μ_j, vectores $\{X_1, X_2, X_3, X_4, \dots, X_k\}$ son independientes y para $\forall j$: $(X_j - \mu_j)^t \Sigma^{-1}(X_j - \mu_j) \longrightarrow \chi^2_{(n_j)}$ tal que $n_j = \dim(E_j)$.

Por teorema, $Y = \hat{Y} \oplus \hat{U}, \hat{Y} \in F = img_{(Imagen)}(F)$ y $\hat{U} \in E$ tal que E es suplemento ortogonal de F en $\mathbb{R}^n$ tal que $\dim_{(Dimensión)}(F) = rg_{(Región)}(X) = k, \dim_{(Dimensión)}(E) = n - k$.

Observación. Supone que hay $k-1$ regresores yregresión tiene término constante tal que la matriz de diseño X tiene k columnas y son linealmente independientes su rango es k. Si regresión no tiene término constante; entonces, el rango de X sería k –

1, $\dim_{(\text{Dimensión})}(E) = n - k + 1$, pero si datos están centrados; entonces, Y está en un espacio vectorial de dimensión $n - 1$ tal que se tendría, de nuevo, $\dim_{(\text{Dimensión})}(E) = n - k$.

2.6 Mínimos cuadrados

El vector β, al igual que la media Xβ, son desconocidos. Si se reemplaza β con un vector $b \in \mathbb{R}^k$, se compara Y respecto a Xb se tiene un vector de residuos que depende de b tal que $Y - Xb$. El estimador de mínimos cuadrados minimiza normal de este vector.

Definición. Según modelo $Y = X\beta + u$, se llama $EMC_{(\text{Estimador de Mínimos Cuadrados})}$ de β a vector $\hat{\beta}$ que minimiza respecto a $\mathbb{R}^k$:

$$SRC_{(\text{Suma de Residuos al Cuadrado})}(b) = \left\|Y - X\hat{\beta}\right\|^2 = \sum_{i=1}^{n}(Y_i - b_1 - b_2x_{i2} - b_3x_{i3} - b_4x_{i4} - \cdots - kx_{ik})^2$$

Tal que:

$$SRC_{(\text{Suma de Residuos al Cuadrado})}(b) = \left\|Y - X\hat{\beta}\right\|^2 = \min_{b \in \mathbb{R}^k}\left\|Y - X\hat{\beta}\right\|^2$$

Cuando $b = \hat{\beta}$ se escribe $SRC_{(\text{Suma de Residuos al Cuadrado})}$ en vez de $SRC_{(\text{Suma de Residuos al Cuadrado})}\left(\hat{\beta}\right)$.

Teorema. Según modelo $Y = X\beta + u$ y matriz X es de rango completo, el estimador de mínimos cuadrados existe, es único y está dado por:

$$\hat{\beta} = (X^tX)^{-1}(X^tY)$$

Demostración:

$$SRC_{(\text{Suma de Residuos al Cuadrado})}(b) = \left\|Y - X\hat{\beta}\right\|^2 = \left[\left(Y^tY\right) - \left(Y^tXb\right) - \left(b^tX^tY\right) + \left(b^tX^tXb\right)\right]$$

$$\frac{\partial SRC_{(\text{Suma de Residuos al Cuadrado})}}{\partial b}(b) = -2X^tY + 2X^tXb \Rightarrow X^tXb = X^tY$$

Pues la matriz X es de rango completo, se tiene que matriz cuadrada X^tX es no singular, la única solución de ecuación matricial anterior sería:

$$\hat{\beta} = (X^t X)^{-1}(X^t Y)$$

Se ha derivado vectorialmente; aunque, es evidente que se puede calcular derivadas parciales respecto de b_j $j = 1,2,3,4, \dots, k$ y se concluye en igual sistema de ecuaciones. **Definición**. Ecuación matricial $X^t Xb = X^t Y$ determina un sistema de k ecuaciones lineales con k incógnitas llamado sistema normal de ecuaciones.

2.7 Máxima verosimilitud

Máxima verosimilitud se basa en ley que siguen las observaciones.

Definición. Si la ley de probabilidad de una variable aleatoria X está dada por una función de densidad tal que f: $\mathbb{R} \longrightarrow \mathbb{R}$ y depende de un parámetro $\theta \epsilon \Theta$, la función de verosimilitud está definida como:

$$\begin{aligned} L_{(x)}: \quad & \Theta \longrightarrow \mathbb{R} \\ : \quad & \theta \longrightarrow L_{(x)}(\theta) = f(x) \end{aligned}$$

La función de densidad depende de x, θ está fijo mientras que función de verosimilitud depende de θ y x permanece fijo. Por otra parte, las hipótesis (N) implican:

$$Y \longrightarrow N_n(X\beta, \sigma^2 I)$$

Tal que, función de verosimilitud sería:

$$L_{(y)}(\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\frac{1}{2\pi\sigma^2} SRC_{(\text{Suma de Residuos al Cuadrado})}(\beta)\right)}$$

Con:

$$SRC_{(\text{Suma de Residuos al Cuadrado})}(\beta) = (Y - X\beta)^t(Y - X\beta) = \|Y - X\beta\|^2$$

Vector aleatorio es un vector formado por una o más variables aleatoria escalar tal que son variables que toman valores en un cuerpo. Normalmente, es de números reales o números complejos:

1. Un vector aleatorio es un vector con componentes equivalentes a variables aleatorias. La ley de un vector aleatorio $X = (X_1, X_2, X_3, X_4, \dots, X_n)^t$ también nombrada ley conjunta de variables $\{X_1, X_2, X_3, X_4, \dots, X_n\}$.

2. Recorderis, si $X = (X_1, X_2, X_3, X_4, \dots, X_n)^t$ es un vector aleatorio de $\mathbb{R}^n$:

$$E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ E(X_3) \\ E(X_4) \\ E(X_5) \\ \vdots \\ E(X_n) \end{bmatrix}, Var(X) = E\left[\left(X - E(X)\right)\left(X - E(X)\right)^t\right]$$

$$= \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) & Cov(X_1, X_4) & \dots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) & Cov(X_2, X_4) & \dots & Cov(X_2, X_n) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) & Cov(X_3, X_4) & \dots & Cov(X_3, X_n) \\ Cov(X_4, X_1) & Cov(X_4, X_2) & Cov(X_2, X_3) & Var(X_4) & \dots & Cov(X_4, X_n) \\ & & \vdots \ \vdots \ \vdots \ \vdots \ \ddots \ \vdots & & & \\ Cov(X_n, X_1) & Cov(X_n, X_2) & Cov(X_n, X_3) & Cov(X_n, X_4) & \dots & Var(X_n) \end{bmatrix}$$

También, matriz $Var(X)$ se escribe como $Cov(X)$ y es llamada matriz de varianzas-covarianzas. En diagonal principal se hallan varianzas y fuera de esta se encuentran covarianzas.

3. Si X es vector aleatorio de $\mathbb{R}^n$ con media μ y varianza Σ. Si A es matriz $(n * m)$, no aleatoria, tal que AX es un vector aleatorio de media $A\mu$ y varianza $A\Sigma A^t$.

4. Toda matriz de varianzas-covarianzas es simétrica semi definida positiva.

5. Dos variables aleatorias X, Y son independientes ssi:

$$\forall x, \forall y \ Pr(X \leq x, Y \leq y) = Pr(X \leq x) Pr(Y \leq y)$$

6. Si X, Y son variables aleatorias independientes, f, g son dos funciones de $\mathbb{R}$ en $\mathbb{R}$. Variables, transformadas, $f(X), g(Y)$ conservan su independencia.

7. Sea Xun vector aleatorio de $\mathbb{R}^n$. Si la función de densidad de X está dada por:

$$f(x) = \left[\left(\frac{1}{(2\pi)^{\left(\frac{n}{2}\right)}\left(\sqrt{|\Sigma|}\right)}\right)\left(e^{\left(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right)}\right)\right] \quad x \in \mathbb{R}^n$$

Tal que $\mu \in \mathbb{R}^n$, Σ es matriz $(n * n)$ simétrica definida positiva, $|\Sigma|$ es determinante de Σ. X sigue una ley normal multivariante de parámetros (μ, Σ) tal que se escribe $X \to N_n(\mu, \Sigma)$.

8. Si $X \to N_n(\mu, \Sigma) \Rightarrow E(X) = \mu$ y $Var(X) = \Sigma$.

9. Sea (X, Y) vectores aleatorios de $\mathbb{R}^n$, si $X \to N_n(\mu_X, \Sigma_X), Y \to N_n(\mu_Y, \Sigma_Y)$ son independientes tal que:

$$(X + Y) \to N_n(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$$

10. Si X es vector aleatorio de distribución $N_n(\mu, \Sigma)$, A es matriz $(m * n)$ tal que $AX \to N_n(A\mu, AA^t\Sigma)$.

11. Sea $X \to N_n(\mu, \Sigma)$, si X se parte en dos sub vectores de m y $n - m$ componentes, respectivamente, si μ y Σ se hacen particiones similares:

$$X = \begin{pmatrix} X^1 \\ X^2 \end{pmatrix}, \mu = \begin{pmatrix} \mu^1 \\ \mu^2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma_{(21)} & \Sigma_{(22)} \end{pmatrix} \Rightarrow X^1 \longrightarrow N_m\left(\mu^1, \Sigma_{(11)}\right)$$

Esta propiedad puede generalizarse a cualquier sub vector. Las leyes marginales de un vector aleatorio de ley normal son normales con sus respectivas medias y varianzas-covarianzas.

12. Con base en estas notaciones e hipótesis, si $Cov(X^1, X^2) = \Sigma_{(12)} = 0$, vectores X^1 y X^2 son independientes. Para sub vectores de un vector de ley normal, no correlación equivale a independencia.

Si X es un vector aleatorio de ley $N_n(\mu, \Sigma)$ y si Σ es no singular tal que $(X - \mu)^t(\Sigma^{-1})(X - \mu) \to \chi_n^2$.

Definición. El $EMV_{(Estimador\ de\ Máxima\ Verosimilitud)}$ de (β, σ^2) es $(\hat{\beta}, \hat{\sigma}^2)$ que maximiza función $L_{(y)}(\beta, \sigma^2)$.

Teorema. Bajo hipótesis N y si matriz X es de rango completo, $EMV_{(Estimador\ de\ Máxima\ Verosimilitud)}$ de β concide con estimador de mínimos cuadrados. Su estimador de $EMV_{(Estimador\ de\ Máxima\ Verosimilitud)}$ respecto a σ^2 sería:

$$\hat{\sigma}^2 = \left(\frac{SRC_{(Suma\ de\ Residuos\ al\ Cuadrado)}}{n}\right)$$

El máximo de función de verosimilitud es:

$$L_{(y)}(\hat{\beta}, \hat{\sigma}^2) = \left(\frac{1}{(2\pi\hat{\sigma}^2)^{\left(\frac{n}{2}\right)}}\right) e^{\left(-\frac{n}{2}\right)}$$

Demostración. La función $\beta \longrightarrow L_{(y)}(\beta, \sigma^2)$ logra su máximo cuando la función $SRC_{(Suma\ de\ Residuos\ al\ Cuadrado)}(\beta)$ alcanza su mínimo tal que el estimador de máxima verosimilitud de β coincide con estimador de mínimos cuadrados de β. También:

$$L_{(y)}(\hat{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\frac{1}{2\sigma^2} SRC_{(Suma\ de\ Residuos\ al\ Cuadrado)}\right)}$$

Tal que para maximizar respecto a σ^2 se toma logaritmo:

$$L_{(y)}(\hat{\beta}, \sigma^2) = \left(-\frac{n}{2}\right) Ln(2\pi) + \left(-\frac{n}{2}\right) Ln(\sigma^2) + \left(-\frac{1}{2\sigma^2} * SRC_{(Suma\ de\ Residuos\ al\ Cuadrado)}\right)$$

Función logaritmo es creciente tal que las dos funciones $L_{(y)}$ y $l_{(y)}$ logran el máximo en mismos valores:

$$\frac{\partial L_{(y)}}{\partial \sigma^2}(\hat{\beta}, \sigma^2) = \left(-\frac{n}{2\sigma^2} + \frac{SRC_{(Suma\ de\ Residuos\ al\ Cuadrado)}}{2\sigma^4}\right) \longrightarrow \hat{\sigma}^2 = \left(\frac{SRC_{(Suma\ de\ Residuos\ al\ Cuadrado)}}{n}\right)$$

Con el fin de comprobar que máximo es suficiente estimar segunda derivada y corroborar su signo:

$$\frac{\partial^2 L_{(y)}}{\partial (\sigma^2)^2}(\hat{\beta}, \sigma^2) = \left(\frac{n}{2\sigma^4}\right) - \left(\frac{SRC_{(Suma\ de\ Residuos\ al\ Cuadrado)}}{2\sigma^6}\right)$$

$$\frac{\partial^2 l_{(y)}}{\partial(\sigma^2)^2}(\hat{\beta},\hat{\sigma}^2) = \left(\frac{n}{2\hat{\sigma}^4}\right) - \left(\frac{n}{\hat{\sigma}^4}\right) = -\left(\frac{n}{2\hat{\sigma}^4}\right)$$

Finalmente:

$$L_{(y)}(\hat{\beta},\hat{\sigma}^2) = \left(\frac{1}{(2\pi\hat{\sigma}^2)^{\left(\frac{n}{2}\right)}}\right) e^{\left(-\frac{n}{2}\right)}$$

2.8 Cambio de origen

La igualdad $\sum_{i=1}^{n} \hat{u}_i = 0$ implica que superficie o hiperplano de regresión pasa por valores medios de variables; es decir, punto $(\bar{x}_2, \bar{x}_3, \bar{x}_4, \dots, \bar{x}_k, \bar{y})$ está sobre superficie de regresión tal que:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \equiv \bar{y} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(\hat{\beta}_1 + \hat{\beta}_2 x_{in} + \cdots + \hat{\beta}_k x_{ik}\right) = \hat{\beta}_1 + \hat{\beta}_2\bar{x}_2 + \cdots + \hat{\beta}_k\bar{x}_k$$

En caso de centrar todas las variables, desvíos respecto a valores medios; es decir, si:

$$y_i^{'} = y_i - \bar{y}, x_{ij}^{'} = x_{ij} - \bar{x}_j \;\; i = 1,2,3,4, \dots, n; j = 1,2,3,4, \dots, k$$

Si se ejecuta la regresión con valores centrados; entonces, $\hat{\beta}_1 = 0$ y la superficie de regresión para por punto $(0,0, \dots,0)$. Entonces, si se cambian estimaciones de otros coeficientes su respuesta sería negativa:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \cdots + \beta_k x_{ik} + u_i \;\; i = 1,2,3,4, \dots, n$$

La regresión con datos originales será:

$$y_i^{'} = \alpha_2 x_{i2}^{'} + \alpha_3 x_{i3}^{'} + \alpha_4 x_{i4}^{'} + \cdots + \alpha_k x_{ik}^{'} + \varepsilon_i \;\; i = 1,2,3,4, \dots, n$$

Tal que la regresión con datos centrados será:

$$y_i^{'} = y_i - \bar{y}, x_{ij}^{'} = x_{ij} - \bar{x}_j \;\; i = 1,2,3,4, \dots, n; j = 1,2,3,4, \dots, k$$

La regresión no tiene término constante pues su estimador es siempre nulo. La relación existente entre coeficientes β_k y α_k indica, en primero lugar, $u_i \approx \varepsilon_i$, pues

ambas son variables aleatorias con ley $N(0,\sigma^2)$. El estudio de relación entre coeficientes es:

$$\begin{aligned} y_i - \bar{y} = \alpha_2(x_{i2} - \bar{x}_2) + \alpha_3(x_{i3} - \bar{x}_3) + \alpha_4(x_{i4} - \bar{x}_4) + \cdots + \alpha_k(x_{ik} - \bar{x}_k) + \varepsilon_i \\ = -(\alpha_2\bar{x}_2 + \alpha_3\bar{x}_3 + \alpha_4\bar{x}_4 + \cdots + \alpha_k\bar{x}_k) \\ + (\alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} \ldots + \alpha_k x_{ik} + \varepsilon_i) \end{aligned}$$

Si se comparan ambas regiones sería:

$$\beta_1 = \bar{y} - (\alpha_2\bar{x}_2 + \alpha_3\bar{x}_3 + \alpha_4\bar{x}_4 + \cdots + \alpha_k\bar{x}_k), \qquad \beta_k = \alpha_k, k = 1,2,3,4, \ldots, j$$

La diferencia única consiste en término constante. Si $\{\hat{\alpha}_k\}$ son estimadores de mínimos cuadrados (MC) de $\{\hat{\alpha}_k\}$:

$$\hat{\beta}_1 = \bar{Y} - (\hat{\alpha}_2\bar{x}_2 + \hat{\alpha}_3\bar{x}_3 + \hat{\alpha}_4\bar{x}_4 + \cdots + \hat{\alpha}_k\bar{x}_k) \quad \forall\, \hat{\beta}_k = \hat{\alpha}_k$$

Observación. Cuando todos los datos han sido centrados:

- Se corre una regresión sin término constante, $\alpha_1 = 0$ tal que el término constante de regresión correspondiente a datos no centrado se estima que el resto de estimadores no cambian:

$$\hat{\beta}_1 = \bar{y} - (\hat{\alpha}_2\bar{x}_2 + \hat{\alpha}_3\bar{x}_3 + \hat{\alpha}_4\bar{x}_4 + \cdots + \hat{\alpha}_k\bar{x}_k)$$

- Matriz de diseño X no contiene columna de unos, $yX^tX = nS_{xx}$ tal que S_{xx} es matriz de varianzas-covarianzas sesgadas de columnas de X. Si S_{xx} es vector de covarianzas sesgadas de columnas de X con vector Y:

$$\hat{\alpha} = (S_{xx}^{-1})(S_{XY}) \Rightarrow \hat{\alpha} = (\hat{\alpha}_2 + \hat{\alpha}_3 + \hat{\alpha}_4 + \cdots + \hat{\alpha}_k)^t$$

- Se puede comprobar:

$$\mathrm{Var}(\hat{\beta}_1) = \left(\frac{\sigma^2}{n}\right)(1 - \bar{x}^t S_{xx}^{-1}\bar{x}) \Rightarrow \bar{x}^t = (\bar{x}_2, \bar{x}_3, \bar{x}_4, , \ldots, \bar{x}_k)$$

2.9 Sumas de cuadrados

Sumas de cuadrados se definen igual que en regresión lineal simple, pero como normas de vectores. **Corolario.** Si $1_n = (1,1,\dots,1)^t \epsilon F = \text{img}_{(\text{Imagen})}(X)$, regresión linal tiene término constante o si $\bar{y} = 0$ tal que:

$$\|Y - 1_n\bar{y}\|^2 = \left\|\hat{Y} - 1_n\bar{y}\right\|^2 + \left\|Y - \hat{Y}\right\|^2 \Rightarrow \bar{y} = \left(\frac{1}{n}\right)\sum y_i$$

Demostración. La proyección ortogonal respecto a $(1_n\bar{y})$ sobre F es $1_n\bar{y}$ tal que por **Teorema**, sea $H_{(\text{Matrix hat o "sombrero"})} = X(X^tX)^{-1}X^t$, vector $\hat{Y} = X\hat{\beta} = \left[H_{(\text{Matrix hat o "sombrero"})}\right]Y$ es proyección ortogonal respecto a producto escalar usual de Y sobre $F = \text{img}_{(\text{Imagen})}(X)$, los vectores $\hat{Y} - 1_n\bar{y}$ y $Y - \hat{Y}$ son ortogonales. La igualdad $\|Y - 1_n\bar{y}\|^2 = \left\|\hat{Y} - 1_n\bar{y}\right\|^2 + \left\|Y - \hat{Y}\right\|^2 \Rightarrow \bar{y} = \left(\frac{1}{n}\right)\sum y_i$ esconsecuencia directa del teorema de Pitágoras.

Las normas anteriores corresponden a sumas de cuadrados que fueron definidas anteriormente:

$$\text{STC}_{(\text{Suma Total de Cuadrados})} = \|Y - 1_n\bar{y}\|^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\text{SEC}_{(\text{Suma de Estimaciones al Cuadrado})} = \left\|\hat{Y} - 1_n\bar{y}\right\|^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$\text{SRC}_{(\text{Suma de Residuos al Cuadrado})} = \left\|Y - \hat{Y}\right\|^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Este corolario indica que si la regresión lineal tiene término constante $(1_n \epsilon F)$ o si $\bar{y} = 0$ tal que

$$\text{STC}_{(\text{Suma Total de Cuadrados})} = \text{SEC}_{(\text{Suma de Estimaciones al Cuadrado})} + \text{SRC}_{(\text{Suma de Residuos al Cuadrado})}$$

Si $1_n \notin F$, regresión no tiene término constante y $\bar{y} \neq 0$, la igualdad anterior no es válida, pero si se tiene:

$$\|Y\|^2 = \left\|\hat{Y}\right\|^2 + \left\|Y - \hat{Y}\right\|^2$$

Ahora bien, si $\bar{y} = 0$:

$$\|Y\|^2 = STC_{(\text{Suma Total de Cuadrados})}, SEC_{(\text{Suma de Estimaciones al Cuadrado})} = \left\|\hat{Y}\right\|^2$$

Teorema. Si $1_n \in F$, regresión tiene término constante, o si $\bar{y} = 0$:

$$SEC_{(\text{Suma de Estimaciones al Cuadrado})} = \hat{Y}^t Y - n\bar{y}^2, SRC_{(\text{Suma de Residuos al Cuadrado})} = Y^t Y - \hat{Y}^t Y, STC_{(\text{Suma Total de Cuadrados})} = Y^t Y - n\bar{y}^2$$

Demostración. La igualdad $STC_{(\text{Suma Total de Cuadrados})} = Y^t Y - n\bar{y}^2$ es conocida. Se demuestra $SRC_{(\text{Suma de Residuos al Cuadrado})} = Y^t Y - \hat{Y}^t Y$:

$$\begin{aligned} SRC_{(\text{Suma de Residuos al Cuadrado})} &= \left\|Y - \hat{Y}\right\|^2 = \left(Y - \hat{Y}\right)^t \left(Y - \hat{Y}\right) = Y^t Y - 2Y^t \hat{Y} + \hat{Y}^t \hat{Y} \\ &= Y^t Y - 2Y^t \hat{Y} + Y^t H^t H Y = Y^t Y - 2Y^t \hat{Y} + Y^t H Y = Y^t Y - 2Y^t \hat{Y} + Y^t \hat{Y} \\ &= Y^t Y - Y^t \hat{Y} \end{aligned}$$

La igualdad $SEC_{(\text{Suma de Estimaciones al Cuadrado})} = \hat{Y}^t Y - n\bar{y}^2$ se puede obtener por diferencia.

Observación. Si no se cumplen condiciones del teorema anterior:

$$SEC_{(\text{Suma de Estimaciones al Cuadrado})} \neq \hat{Y}^t Y - n\bar{y}^2$$

Tal que no se puede obtener por diferencia de las otras sumas de cuadrados. Resultados como este se sueles estimar y escribir en una tabla ANOVA, presentada en prueba de significancia de regresión.

2.10 Estimación σ^2

Teorema. Según modelo $Y = XB + U$ con $E(U) = 0$, $Var(U) = \sigma^2 I$ y si matriz X es de rango completo k, no aleatoria:

$$S_R^2 = \left(\frac{SRC_{(\text{Suma de Residuos al Cuadrado})}}{n - k}\right) \Rightarrow \text{estimador de } \sigma^2$$

Demostración. $\hat{U} = (I - H)Y$ con $H = X(X^tX)^{-1}X^t$. Se demuestra fácilmente que la matriz $(I - H)$ es simétrica e idempotente. Además:

$$(I - H)XB = 0$$

Si se toma en cuenta que $tr(AB) = tr(BA)$:

$$tr(H) = tr[X(X^tX)^{-1}X^t] = tr(X(X^tX)^{-1}X^t) = tr\left(I_{(k*h)}\right) = k$$

$$\begin{aligned}\hat{U}^t\hat{U} = Y^t(I - H)Y &= (X\beta + U)^t(I - H)(X\beta + U) \\ &= \beta^tX^t(I - H)X\beta + \beta^tX^t(I - H)U + U^t(I - H)X\beta + U^t(I - H)U \\ &= U^t(I - H)U\end{aligned}$$

$$SRC_{(\text{Suma de Residuos al Cuadrado})} = tr\left(\hat{U}^t\hat{U}\right) = tr\left(\hat{U}^t(I - H)U\right) = tr\left((I - H)U\hat{U}^t\right)$$

$$\begin{aligned}E\left(SRC_{(\text{Suma de Residuos al Cuadrado})}\right) &= E\left[tr\left((I - H)U\hat{U}^t\right)\right] = tr\left[(I - H)E\left(U\hat{U}^t\right)\right] \\ &= \sigma^2 tr[(I - H)] = \sigma^2(n - k)\end{aligned}$$

Repaso a Álgebra Lineal:

18. Vectores $(u_1, u_2, u_3, u_4, \dots, u_k)$ de $\mathbb{R}^n$ son linealmente independientes ssi:

$$a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4 + \cdots + a_ku_k = 0$$

Implica que todos los coeficientes $a_j = 0$ o nulos. Si no, son linealmente independientes. Un conjunto infinito B de vectores es linealmente independiente ssi todo conjunto B de vectores es linealmente independiente ssi todo sub conjunto finito de B es linealmente independiente.

19. Si vectores $\{u_j\}$ son linealmente dependientes existe, en igualdad anterior, un coeficiente $a_j \neq 0$ o no nulo tal que el vector correspondiente a dicho coeficiente es combinación lineal del resto de vectores. Por ejemplo: $a_j \neq 0 \Rightarrow u_k = \left(-\frac{1}{a_k}\right)(a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4 + \cdots + a_{k-1}u_{k-1})$.

20. Bases y dimensión. Sea E un espacio vectorial y B un sub conjunto de E:

    d) Si B genera a E ssi ∀ vector de E es una combinación lineal de vectores de B.

    e) Si B es linealmente independiente se dice que B es una base de E.

    f) Si B es una base de E, el número de elementos de B se llama dimensión de B, si B es infinito se dice que E es dimensión infinita.

21. Número de filas linealmente independientes, en una matriz $A_{(n*m)}$ es igual a número de columnas linealmente independientes.

22. Sea A una matriz $A_{(n*m)}$. El número de filas o columnas linealmente independientes se llama rango de matriz y se denota como $rg(A_{(Rango)})$.

23. De bases y dimensión. Sea E un espacio vectorial y B un sub conjunto de E, es obvio que $rg(A_{(Rango)}) \leq \min\{n, m\}$. Si se logra igualdad $[rg(A_{(Rango)}) \leq \min\{n, m\}]$ se dice que matriz A es rango completo.

24. Matriz cuadrada A es simétrica ssi al cambiar filas por columnas la matriz A no cambia; es decir, ssi $A^t = A$ donde A^t es transpuesta de A.

25. Matriz cuadrada A es diagonal ssi es cuadrada y todos los elementos fuera de la diagonal principal son 0 o nulos:

$$A = \text{diag}_{(\text{Diagonal})}\{a_1 + a_2 + a_3 + a_4 + \cdots + a_n\} = \begin{bmatrix} a_1 & 0 & 0 & 0 & 0 & & 0 \\ 0 & a_2 & 0 & 0 & 0 & & 0 \\ 0 & 0 & a_3 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & a_4 & 0 & & 0 \\ 0 & 0 & 0 & 0 & a_5 & & 0 \\ & & \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & a_n \end{bmatrix}$$

26. Matriz identidad es matriz diagonal con todos $a_i = 0$:

$$I = \text{diag}_{(\text{Diagonal})}\{1,1,1,1, \dots ,1\}$$

Cuando requiere indicar el orden de matriz se estribe I_n en vez de I.

27. Si A matriz cuadrada $(n * n)$ es no singular ssi existe una matriz B cuadrada $(n * n)$:

$$AB = BA = I$$

Matriz B se llama inversa de A y se escribe A^{-1}

28. $(AB)^t = B^t A^t$. Si matrices (A, B) son no singulares y de igual dimensión $(AB)^{-1} = B^{-1}A^{-1}$.

29. Una matriz cuadrada $A_{(n*n)}$ es simétrica semi definida positiva, escrita como $A_{(n*n)} \geq 0$, ssi es simétrica y $\forall x, x \in \mathbb{R}^n, x^t Ax \geq 0$ se dice simétrica definida positiva, escrita $A > 0$, ssi es se mi definida positiva y $x^t Ax = 0 \Rightarrow x = 0$.

30. Sea E, F dos espacios vectoriales sobre un mismo cuerpo de escalares K. Una función $f: E \longrightarrow F$ se dice aplicación lineal ssi $\forall a \in k, \forall x, y \in E$ se tiene que $f(ax + y) = af(x) + f(y)$.

31. $\forall A_{(m*n)}$ define una aplicación lineal $\mathbb{R}^n$ en $\mathbb{R}^m$:

$$A: \mathbb{R}^n \rightarrow \mathbb{R}^m: x \rightarrow Ax$$

A su vez toda aplicación lineal de $\mathbb{R}^n$ en $\mathbb{R}^m$ define una matriz $(m * n)$.

32. Si A es aplicación lineal de $\mathbb{R}^n$ en $\mathbb{R}^m$ se llama espacio imagen de A al conjunto:

$$\text{img}_{(\text{Imagen})}(A) = \{w \in \mathbb{R}^m | w = Au, u \in \mathbb{R}^n\}$$

Se llama núcleo de A al conjunto:

$$\text{Ker}(A) = \{u \in \mathbb{R}^n | Au = 0\}$$

Se puede demostrar que $\dim_{(\text{Dimensión})}\left[\text{img}_{(\text{Imagen})}(A)\right] = \text{rg}\left(A_{(\text{Rango})}\right)$ y $\dim_{(\text{Dimensión})}\left[\text{img}_{(\text{Imagen})}(A)\right] + \dim_{(\text{Dimensión})}\left[\text{Ker}_{(\text{Kernel o núcleo})}(A)\right] = n$

33. Si $A_{(n*n)}$ es matriz cuadrada $(n * n)$ simétrica definida positiva, se define el producto escalar de dos vectores u, v de $\mathbb{R}^n$ respecto de A por $\langle u, u\rangle_{(A)} = (u^t)(Au)$ se define norma de v respecto de A por $\|u\|_{(A)} = \sqrt{(u^t)(Au)}$. Si A es matriz identidad no hace falta sub índices tal que se dice que son producto escalar y norma usuales.

34. Si A es matriz cuadrada $(n * n)$ se llama traza de A a suma de elementos de su diagonal principal tal que $\text{tr}_{(\text{Traza de A})}(A) = \sum_{i=1}^{n}(a_{ii})$.

Sea A matriz $(n * m)$, B matriz $(m * n)$ tal que $\text{tr}_{(\text{Traza de AB})}(AB) = \text{tr}_{(\text{Traza de BA})}(BA)$.

2.10.1 Propiedades

Teorema. Si matriz X es de rango completo, no aleatoria y $E(U) = 0$:

$$\hat{\beta} = (X^tX)^{-1}X^tY \Rightarrow \text{Estimador insesgado de } \beta$$

Demostración. X es no estocástica, $E(Y) = X\beta$ tal que:

$$E\left(\hat{\beta}\right) = E[(X^tX)^{-1}X^tY] = (X^tX)^{-1}X^tE(Y) = (X^tX)^{-1}X^tX\beta = \beta$$

Teorema. Si matriz X es de rango completo, no aleatoria y $\text{Var}(U) = (\sigma^2)(I)$:

$$\mathrm{Var}(\hat{\beta}) = (\sigma^2)(X^tX)^{-1}$$

Demostración. Si A es matriz de constantes y Y es vector aleatorio $\mathrm{Var}(AY) = (A)\mathrm{Var}(Y)(A^t)$:

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}[(X^tX)^{-1}X^tY] = [(X^tX)^{-1}X^t]\mathrm{Var}(Y)[(X^tX)^{-1}X]$$
$$= (\sigma^2)[(X^tX)^{-1}X^t][(X^tX)^{-1}X] = (\sigma^2)[(X^tX)^{-1}]$$

Corolario. Si $E(U) = 0, \mathrm{Var}(U) = (\sigma^2)(I)$ y matriz X es de rango completo k; entonces, $\widehat{\mathrm{Var}}(\hat{\beta}) = S_R^2(X^tX)^{-1}$ tal que $S_R^2(X^tX)^{-1} = \left(\frac{\mathrm{SRC}_{(\text{Suma de Residuos al Cuadrado})}}{n-k}\right) \Rightarrow$ Estimador insesgado de $\mathrm{Var}(\hat{\beta})$.

Un estimador $\hat{\theta}$ de θ es lineal si y solo si es combinación lineal de observaciones tal que si $\theta \in \mathbb{R}^k$ y vector de observaciones $Y \in \mathbb{R}^k$, $\hat{\theta}$ es lineal y solo si $\in A_{(\mathrm{Matriz:k*n})} \Rightarrow \hat{\theta} = AY$.

Definición. Si $\hat{\theta}$ y $\tilde{\theta}$ son estimadores insesgados de θ tal que $\theta \in \mathbb{R}$ se dice que $\hat{\theta}$ es mejor que $\tilde{\theta}$ ssi:

$$\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})$$

Observación. Si $\hat{\theta}$ y $\tilde{\theta}$ son vectores de $\mathbb{R}^k \Rightarrow \mathrm{Var}(\hat{\theta})$ y $\mathrm{Var}(\tilde{\theta})$ son matrices $k * k$. Esta desigualdad no se aplica directamente; aunque, se puede extender su significado tal que matriz $\mathrm{Var}(\hat{\theta}) - \mathrm{Var}(\tilde{\theta})$ es simétrica semi definida positiva; es decir, $\forall\, a \in \mathbb{R}$:

$$(a^t)(a)\mathrm{Var}(\tilde{\theta}) \geq (a^t)(a)\mathrm{Var}(\hat{\theta})$$

2.11 Errores no normales

Resultados teorema, $Y = \hat{Y} \oplus \hat{U}, \hat{Y} \in F = \mathrm{img}_{(\mathrm{Imagen})}(F)$ y $\hat{U} \in E$ tal que E es suplemento ortogonal de F en $\mathbb{R}^n$ tal que $\dim_{(\mathrm{Dimensión})}(F) = \mathrm{rg}_{(\mathrm{Región})}(X) = k, \dim_{(\mathrm{Dimensión})}(E) = n - k$ se fundamentan en supuesto que errores siguen una ley normal, pero el cálculo de intervalos de confianza y contrastes de hipótesis se basan en el mismo. Sin embargo, ¿qué sucede si errores no se ajustan a ley normal? Una de las consecuencias es que estimadores de mínimos cuadrados no coinciden con estimadores de máxima

verosimilitud y, más importante, pruebas e intervalos de confianza podrían no tener validez.

CAPÍTULO III

CAPÍTULO III

INFERENCIA PARAMÉTRICA

La inferencia es intervalos, regiones, de confianza y pruebas de hipótesis se pueden hacer para un parámetro en particular, un grupo de parámetros o sus combinaciones lineales.

3.1 Pruebas de hipótesis e intervalos de confianza en parámetros individuales

Es una generalización de resultados obtenidos para regresión lineal simple y son basados en hipótesis de normalidad de errores, pero se pueden extender al caso no normal a condición de tener suficientes observaciones que h_{ii} no sean muy grandes, mayores que $3\left(\frac{k}{n}\right)$.

De $\hat{\beta} \longrightarrow N_k[\beta, \sigma^2(X^tX)^{-1}]$ y propiedad $X \longrightarrow N_n(\mu, \Sigma)$ si X se parte en dos sub vectores de m y $n - m$ componentes, respectivamente, si μ y Σ se hacen particiones similares tal que $X = \begin{pmatrix} X^1 \\ X^2 \end{pmatrix}, \mu = \begin{pmatrix} \mu^1 \\ \mu^2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma_{(21)} & \Sigma_{(22)} \end{pmatrix}$ tal que $X^1 \longrightarrow N_m(\mu^1, \Sigma_{(11)})$ aunque se puede generalizar a cualquier sub vector e incluso las leyes marginales de un vector aletorio de ley normal son normales con sus respectivas medias y varianzas-covarianzas, $\hat{\beta}$ y $\hat{U}$ son normales; además, por propiedad, en que notaciones e hipótesis anteriores, si $\mathrm{Cov}(X^1, X^2) = \Sigma_{(12)} = 0$ los vectores X^1 y X^2 son independientes tal que para sub vectores de un vector de ley normal la no correlación equivale a independencia, se inducir que para $\forall_j\ \hat{\beta}_j \rightarrow N(\beta_j, \sigma^2 u_{jj})$ tal que u_{jj} es j-ésimo elemento de diagonal principal de matriz $(X^tX)^{-1}$. De $\hat{\beta}$ y $\hat{U}$ son independientes se infiere que $\hat{\beta}_j$ y s^2 son independientes.

A su vez, estas propiedades permiten demostrar el teorema. Con hipótesis N, si X es de rango completo y no aleatorio:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{s_R^2 u_{jj}}} \to t_{(n-k)}$$

Definición. La cantidad $ee_{(\text{error estándar})}(\hat{\beta}_j) = \sqrt{s^2 u_{jj}}$.

El cociente:

$$t_{j\ (\text{Razón t de Student})} = \frac{\hat{\beta}_j}{\sqrt{s_R^2 u_{jj}}}$$

Corolario. Con hipótesis N, si X es de rango completo y no aleatorio para $\forall_j$:

$$\hat{\beta}_j \pm \sqrt{s_R^2 u_{jj}} t_{\left(n-k;\frac{\alpha}{2}\right)}$$

Es intervalo de confianza para β_j de nivel $\eta = 1 - \alpha$. $t_{\left(n-k;\frac{\alpha}{2}\right)}$ es fractil de orden $\left(1-\frac{\alpha}{2}\right)$tal que $\left(\frac{\alpha}{2}\right)$ es área de cola derecha de ley t de Student con $n - k_{(\text{grados de libertad})}$.

Corolario. Con hipótesis N, si X es de rango completo y no aleatorio la prueba de hipótesis es de nivel α. Si una prueba de hipótesis es de nivel α indica que probabilidad de error de primer tipo es α. "la probabilidad α recibe el nombre de nivel de significancia o, en forma más sencilla, nivel de prueba.

Aun cuando se recomiendan con frecuencia pequeños valores de α, el valor real α para usar en un análisis es un tanto arbitrario. no aceptar "rechazar" o no rechazar "aceptar" una hipótesis nula depende de α, nivel de significancia o probabilidad de cometer error tipo I, probabilidad de rechazar la hipótesis cuando es verdadera. α se fija en niveles de 1, 5 o cuando mucho, 10%, pero no hay nada "sagrado" acerca de estos valores tal que cualquier otro valor sería por igual apropiado.

No obstante, comúnmente se fijan estos niveles de α para control de calidad, diseños experimentales o investigaciones sociales, respectivamente y, de igual manera, los econometristas tienen por costumbre fijar el valor α en estos niveles como máximo y escogen un estadístico de prueba que haga la probabilidad de comer un error tipo II sea lo más pequeño posible.

Como uno menos la probabilidad de cometer error tipo II se conoce como **potencia de prueba**, este procedimiento equivale a maximizar la potencia de prueba. No se acepta $H_0: \beta_j = 0$ vs H_a o $H_1: \beta_j \neq 0$ a nivel α si y sólo si $|t_{j\ (\text{Razón t de Student})}| > t_{\left(n-k;\frac{\alpha}{2}\right)}$.

H_0 se lee "H subíndice 0", pues H significa hipótesis y el subíndice 0 señala "no hay diferencia".

En cambio, H_1 o H_a, leído como "H subíndice 1 o a" señala que "al menos 1 es diferente respecto al resto de los tratamientos". La demostración matemática que lo sustenta se encuentra en la siguiente nota al pie y da razón que exista H_0 e H_a en una prueba de dos colas.

Además, con base en una muestra de una variable aleatoria X de ley P_θ en decidir entre dos hipótesis:

$H_0: \theta \in \theta_{0\ (\text{hipótesis nula o privilegiada})}$ vs $H_1: \theta \in \theta_{1\ (\text{hipótesis alternativa o alterna})}$ tal que $\theta_0 \cup \theta_1 = \theta$ y $\theta_0 \cap \theta_1 = \emptyset$. Si d_0 y d_1 representan, respectivamente, las decisiones de no rechazar H_0 o H_1 y $D = \{d_0, d_1\}$.

Entonces: sea $X: \Omega \longrightarrow E \subseteq \mathbb{R}$ una variable aleatoria de ley P_θ. Se llama "Prueba de Hipótesis Pura" o test puro a toda aplicación: $\emptyset: E^n \longrightarrow D$ tal que $\emptyset$ equivale a particionar E^n en dos conjuntos: $W = \emptyset^{-1}(d_0) = \{x \in E^n: \text{si se observa x, no se acepta } H_0\}$ y $W^c = \emptyset^{-1}(d_1) = \{x \in E^n: \text{si se observa x, no se rechaza } H_0\}$.

El nombre de H_0 proviene que H_0 se asumirá como verdadera, salvo que datos muestrales indiquen su falsedad. Nula debe entenderse como neutra. **H_0 nunca se**

considera probada o demostrada, salvo estudiando todos los datos de la población, puede diferir en un valor pequeño imperceptible para el muestreo, **que puede ser imposible; aunque, puede no ser aceptada por los datos**.

Con base en esto, **no se debe afirmar "se acepta $\mathbf{H_0}$"**, siendo lo correcto **"no se rechaza $\mathbf{H_0}$"** y, por abuso de lenguaje, es común hallar la expresión "se acepta la H_0" en lugar de "no se rechaza H_0". Generalmente, H_0 se elige según con el principio de simplicidad científica, que establece que únicamente se abandona un modelo simple a favor de otro más complejo cuando la evidencia a favor de este último sea fuerte.

Por el carácter dicotómico, si no se acepta H_0, automáticamente no se rechaza H_1. Finalmente, se llama **riesgo de primera especie a la probabilidad de rechazar $\mathbf{H_0}$ cuando es verdadera**: $\alpha_0(\emptyset) = P_0(W_{(\text{Región critica de prueba})})$, mientras que el **riesgo de segunda especie es la probabilidad de aceptar $\mathbf{H_0}$ cuando es falsa**: $\beta_\theta(\emptyset) = P_1\left(W^c_{(\text{Región de aceptacion})}\right) = 1 - P_1(W)$.

Se llama "Potencia de una prueba" a la probabilidad de no aceptar H_0 cuando es falsa $-P_1(W) = 1 - \beta_\theta(\emptyset) -$ y, también, se denomina "Nivel de significación de una prueba" al valor $\alpha = \sup \alpha_\theta(\emptyset)\ \theta \in \theta_0$.

Cuando $\theta \subseteq \mathbb{R}$ se estudian problemas del tipo: $P_0: \begin{cases} H_0: \theta = \theta_0, \\ H_1: \theta = \theta_1 \text{ con } \theta_0 \neq \theta_1 \end{cases}$,

$P_1: \begin{cases} H_0: \theta \leq \theta_0, \\ H_1: \theta > \theta_{0\ (\text{Prueba de cola derecha})} \end{cases}$, $P_2: \begin{cases} H_0: \theta \geq \theta_0, \\ H_1: \theta < \theta_{0\ (\text{Prueba de cola izquierda})} \end{cases}$,

$P_3: \begin{cases} H_0: \theta_0 \leq \theta \leq \theta_1, \\ H_1: \theta < \theta_{0\ (\text{Prueba de cola izquierda})} \text{ o } \theta > \theta_{1\ (\text{Prueba de cola derecha})} (\theta_0 < \theta_1) \end{cases}$,

$P_4: \begin{cases} H_0: \theta < \theta_0 \text{ o } \theta > \theta_1, \\ H_1: \theta_0 \leq \theta \leq \theta_1, \end{cases}$ o $P_5: \begin{cases} H_0: \theta = \theta_0, \\ H_1: \theta \neq \theta_1 \end{cases}$.

El **teorema Neyman-Pearson** indica, en contexto de prueba propuesta: $\forall\ \alpha \in [0,1], \epsilon$ un test puro $\emptyset$, de nivel α, de potencia máxima, definido por la región critica: $W = \left\{x \in E^n : \frac{L(\theta_0, x)}{L(\theta_1, x)} \leq k\right\} \longrightarrow k$ se determina de la condición $\alpha = P_0(W)$.

L representa la función de verosimilitud y x una observación de la muestra. Otras pruebas importantes, de nivel α, son:

A): $\begin{cases} H_0: \mu > \mu_0, \\ H_1: \mu \leq \mu_0 \end{cases}$ y

B): $\begin{cases} H_0: \mu < \mu_0, \\ H_1: \mu \geq \mu_0 \end{cases}$.

Si $\chi^2_{\text{Calculada}}$, $T_{\text{Calculada}}$ (T value) o $F_{\text{Calculada}}$ (F value) es menor que χ^2_{Tablas}, T_{Tablas} de Student o F_{Tablas} de Fisher Tablas $(\alpha_{(\text{Tamaño de error o nivel de significancia})}, k-1_{(\text{Grados de libertad del numerador})}, N-k_{(\text{Grados de libertad del denominador})})$ se tiene el **criterio más formal y/o riguroso de lectura de una ANOVA** que no se rechaza hipótesis nula, que indica igualdad de medias de tratamientos ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \cdots = \mu_k = \mu\,(0)$ o $H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = \cdots \tau_k = 0$) y no se acepta la alternativa, que indica diferencia de al menos una media de tratamiento respecto al resto ($H_a: \mu_i \neq \mu_j$para algún $i \neq j$ o $H_a: \tau_i \neq 0$ para algún i).

Un conjunto generador se representa a partir que vectores $\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4, \ldots, \upsilon_n$ de un espacio vectorial V genera a V si todo vector en V se puede escribir como una combinación lineal de los mismos; es decir, $\forall\, \upsilon \in V_{(\text{Espacio Vectorial})}$ ε escalares $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \ldots, \alpha_n \Rightarrow \upsilon = \alpha_1\upsilon_1 + \alpha_2\upsilon_2 + \alpha_3\upsilon_3 + \alpha_4\upsilon_4 \ldots + \alpha_n\upsilon_n$.

Además, un espacio generado por un conjunto de vectores se genera a partir que sea $\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4, \ldots, \upsilon_k$ k vectores de un espacio vectorial V.

El espacio generado por $\{\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4, \ldots, \upsilon_k\}$ es el conjunto de combinaciones lineales $\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4, \ldots, \upsilon_k$; es decir, $\text{gen}\{\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4, \ldots, \upsilon_k\} = \{\upsilon | \upsilon = \alpha_1\upsilon_1 + \alpha_2\upsilon_2 + \alpha_3\upsilon_3 + \alpha_4\upsilon_4 \ldots + \alpha_k\upsilon_k\}$.

Por lo tanto, se dice que es **linealmente dependiente** (L. D. o D. L. en inglés) si sean $\alpha_1\upsilon_1 + \alpha_2\upsilon_2 + \alpha_3\upsilon_3 + \alpha_4\upsilon_4 \ldots + \alpha_n\upsilon_n$, n vectores en espacio vectorial V tal que los vectores son linealmente dependientes si existen n escalares $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \ldots, \alpha_n$ NO TODOS SON IGUALES A CERO; es decir, con algún $\alpha_i \neq 0 \Rightarrow \alpha_1\upsilon_1 + \alpha_2\upsilon_2 + \alpha_3\upsilon_3 +$

$\alpha_4 \upsilon_4 \ldots + \alpha_n \upsilon_n = \vec{0}$. Caso contrario, estos serán **linealmente independientes** (L. I. o I. L. en inglés) tal que, la única combinación lineal que da cero en la TRIVIAL es $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \cdots = \alpha_n = 0$.

El nivel crítico o valor ρ de esta prueba es:

$$\hat{\alpha} = 2\Pr\left(T_{(n-k):(\text{Variable aleatoria de ley } t_{(n-k)})} \middle| \middle| t_{j\ (\text{Razón t de Student})} \middle|\right)$$

Pr$(>$ **F**$)$ **o** $\boldsymbol{\rho - }$ **Value**: Es la probabilidad de observar un valor muestral tan extremo o más extremo que el valor observado dado que H_0 es verdadera o es una manera de expresar la probabilidad H_0 no sea verdadera.

Complementariamente, Pr$(>$ F$)$ es el nivel de probabilidad que H_a caiga en la zona de rechazo o, complementariamente, no hay ninguna diferencia entre los dos grupos o no hay correlación entre un par de características.

Según (Gujarati & Porter, 2010), $\rho -$ Value, valor de probabilidad, nivel observado, exacto de significancia o probabilidad exacta de cometer error tipo I, estima la probabilidad real de obtener un valor del estadístico de prueba tan grande o mayor que el obtenido en un ejercicio o, en otras palabras, es el nivel de significancia más bajo al cual puede rechazarse una hipótesis nula, mientras que Leek y Peng afirman que Pr$(>$ F$)$ o $\rho -$ Value es la cima del iceberg.

De acuerdo con la publicación "Estadísticos emiten alerta sobre mal uso de Pr$(>$ F$)$ o $\rho -$ Value (www. nature. com/news/statisticians $-$ issue $-$ warning $-$ over $-$ misuse $-$ of $-$ p $-$ values $-$ 1.19503#/ref $-$ link $-$ 1) Wasserstein & Lazar (Wasserstein, R. L. & Lazar, N. A. advance online publication The American Statistician, 2016) "el valor de P es una prueba común para juzgar la fuerza de la evidencia científica, contribuye al número de resultados de la investigación que no puede ser reproducido, la asociación estadística americana (ASA) advierte en un comunicado publicado hoy.

El grupo ha tomado el paso inusual de emitir principios para guiar el uso del $\rho -$ Value, que dice **no se puede determinar si una hipótesis es verdadera o si los resultados son importantes**. En su comunicado, la **ASA informa a investigadores evitar conclusiones científicas o tomar decisiones de política basadas solo en valores ρ**. Los investigadores **deben describir no sólo el análisis de datos que produjo resultados estadísticamente significativos**, la sociedad dice, **pero todas pruebas y decisiones en cálculos**.

De lo contrario, pueden parecer falsamente robustos resultados. El más pequeño el valor de $\rho -$ Value indica menos probabilidad que un conjunto observado de valores ocurra por casualidad, suponiendo que la hipótesis nula es verdadera. **Un valor de $\rho \leq 0.05$ se toma**, generalmente, **para expresar que un resultado es estadísticamente significativo y garantiza la publicación** o **hay 95% de probabilidad que una hipótesis dada es correcta. Pero no es necesariamente verdadero**, según notas de instrucción de ASA.

En cambio, **significa si la hipótesis nula** (H_0) **es verdadera y todos los otros supuestos son válidos,** tal que hay un 5% de probabilidad de obtener resultados al menos tan extrema como el observado. **Un $\rho -$ Value no puede indicar la importancia de un resultado.** Vickers dice que **"investigadores deben tratar la estadística como una ciencia y no una receta"**.

3.2 Pruebas de hipótesis afines

Es importante probar si un sub vector de parámetros es nulo, hipótesis lineal, o más, generalmente, si una combinación lineal de parámetros tiene un valor particular, hipótesis afín. **Definición**. Sean $C_{(\text{Matriz q}*\text{k de rango q} < \text{k})}, \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}$. Hipótesis de forma $H_0: C\beta = \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}$ se llama hipótesis afín tal que si $\psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}} = 0$ se nombra hipótesis lineal. Su objetivo es probar, a un nivel de significancia dado

(α), hipótesis $H_0: C\beta = \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}$ vs $H_0: C\beta \neq \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}$. Existen dos procedimientos:

3.2.1 A partir de matriz de varianzas-covarianzas.

Por teorema de Gauss-Markov el mejor estimador lineal insesgado de $\psi = C\beta$ es $\widehat{\psi} = C\hat{\beta}_{(\text{estimador de mínimos cuadrados ordinarios}-\text{MCO}-\text{de } \beta)}$.

Teorema. Con hipótesis usuales de mínimos cuadrados ordinarios (MCO) tal que $C_{(\text{Matriz q}*\text{k de rango q} < \text{k})}, \text{rg}_{(\text{Rango})}(X) = k$:

a) $\widehat{\psi} \to N_q(\psi, \sigma^2 S), S = CC^t(X^t X)^{-1}$.

   Demostración. Este inciso es consecuencia de propiedad $\hat{\beta} \longrightarrow N_k[\beta, \sigma^2(X^t X)^{-1}]$ y si X es un vector aleatorio de distribución $N_n(\mu, \Sigma)$, $A_{(\text{Matriz m}*\text{n})}$ tal que $AX \longrightarrow N_m[A\mu, A\Sigma A^t]$.

b) Matriz S es no singular tal que $\left[\left(\frac{1}{\sigma^2}\right)(\widehat{\psi} - \psi)^t S^{-1}(\widehat{\psi} - \psi) \to \chi^2_q\right]$.

   Demostración. Este inciso es consecuencia de propiedad $\hat{\beta} \longrightarrow N_k[\beta, \sigma^2(X^t X)^{-1}]$ y si X es un vector aleatorio de distribución $N_n(\mu, \Sigma)$, Σ no singular tal que $(X - \mu)^t \Sigma^{-1}(X - \mu) \longrightarrow \chi^2_n$.

c) $F_{(\psi)} = \left[\frac{(\widehat{\psi}-\psi)^t S^{-1}(\widehat{\psi}-\psi)}{(q)(S^2_R)}\right] \to F_{(q,n-k)}$.

   Demostración. Este inciso es consecuencia de independencia de $\widehat{\psi}, s^2_R$ y tanto numerador como denominador siguen leyes $\chi^2_{(\text{Independientes})}$.

Lema. Si C es rango completo, el estimador de mínimos cuadrados bajo hipótesis $H_0: C\beta = \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}$ está dado por $\hat{\beta}^0 = \hat{\beta}_{(\text{estimador de mínimos cuadrados sin restricción})} - C^t(X^t X)^{-1} S^{-1}(\widehat{\psi} - \psi_0)$ tal que $\widehat{\psi} = C\hat{\beta}, S = CC^t(X^t X)^{-1}$.

Demostración. Se minimiza $\|Y - Xb\|^2$ sujeto a restricción $Cb = \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}$ en que se emplean multiplicadores de Lagrange:

$$f\left(b,\lambda_{(\text{Letra griega lambda})}\right)=\|Y-Xb\|^2+2\lambda^t\left(Cb-\psi_{0\,(\text{Vector fijo de }\mathbb{R}^q)}\right)$$

$$\Rightarrow\frac{\partial(f)}{\partial(b)}\left(b,\lambda_{(\text{Letra griega lambda})}\right)=-2X^tY+2X^tb+2C^t\lambda$$

$$\Rightarrow\frac{\partial(f)}{\partial(b)}\left(b,\lambda_{(\text{Letra griega lambda})}\right)=Cb-\psi_{0\,(\text{Vector fijo de }\mathbb{R}^q)}$$

Está comprendida en "Optimización con restricciones". El "Método multiplicadores de Lagrange" consiste en construir una función $W[x,y]$ formada por suma de función $Z[x,y]$ más restricción $\varphi_{(\text{Letra griega minúscuma phi})}[x,y]$ multiplicada por $\lambda_{(\text{Letra griega minúscuma lambda})}$ tal que hay que hallar puntos extremos de $W[x,y]$. Este teorema es instrumento teórico más clásico para resolver problemas de optimización con restricciones.

El planteamiento de un problema de optimización. Problema de programación matemática es optimizar $f(X_1,X_2,X_3,X_4,\dots,X_n)$ sujeto a restricciones (sar) $(X_1,X_2,X_3,X_4,\dots,X_n)\in F$ tal que $f\colon A\subseteq\mathbb{R}^n\longrightarrow\mathbb{R}, y=f(X_1,X_2,X_3,X_4,\dots,X_n)$ es función objetivo. $(X_1,X_2,X_3,X_4,\dots,X_n)$ son variables de decisión. $F\subseteq\mathbb{R}^n$ es conjunto de oportunidades o región factible. $F\cap X, A=\text{Dom}_{(\text{Dominio})}(f)$ tal que X es conjunto dado por restricciones del problema.

En "Método de multiplicadores de Lagrange", la solución (x^*,y^*) del problema (ρ) verifica:

$$\left[\frac{\left(\frac{\partial f}{\partial x}\right)}{\left(\frac{\partial f}{\partial y}\right)}\right]=\left[\frac{\left(\frac{\partial g}{\partial x}\right)}{\left(\frac{\partial g}{\partial y}\right)}\right]\Rightarrow\exists\,\lambda^*_{(\text{Letra griega minúscula lambda})}:\left[\frac{\left(\frac{\partial f}{\partial x}\right)}{\left(\frac{\partial g}{\partial x}\right)}\right]=\left[\frac{\left(\frac{\partial f}{\partial y}\right)}{\left(\frac{\partial g}{\partial y}\right)}\right]$$

$$=\lambda^*_{(\text{Letra griega minúscula lambda})}\Rightarrow:$$

$$\begin{cases}\left(\frac{\partial f}{\partial x}\right)=\left(\lambda^*_{(\text{Letra griega minúscula lambda})}\right)\left(\frac{\partial g}{\partial x}\right)\\ \left(\frac{\partial f}{\partial y}\right)=\left(\lambda^*_{(\text{Letra griega minúscula lambda})}\right)\left(\frac{\partial g}{\partial y}\right)\end{cases}\Rightarrow\quad\begin{cases}\left(\frac{\partial f}{\partial x}\right)(x^*,y^*)-\left(\lambda^*_{(\text{Letra griega minúscula lambda})}\right)\left(\frac{\partial g}{\partial x}\right)(x^*,y^*)=0\\ \left(\frac{\partial f}{\partial y}\right)(x^*,y^*)-\left(\lambda^*_{(\text{Letra griega minúscula lambda})}\right)\left(\frac{\partial g}{\partial y}\right)(x^*,y^*)=0\end{cases}\quad(I)$$

Se llama $L(x,y,\lambda)=f(x,y)-\lambda[g(x,y)-b]$, (I) y restricción de (ρ) se escribe:

$$\begin{cases} \left(\dfrac{\partial L}{\partial x}(x^*, y^*, \lambda^*)\right) = 0 \quad (1) \\ \left(\dfrac{\partial L}{\partial y}(x^*, y^*, \lambda^*)\right) = 0 \quad (2) \quad (II) \\ \left(\dfrac{\partial L}{\partial \lambda}(x^*, y^*, \lambda^*)\right) = 0 \quad (3) \quad \Leftrightarrow -[g(x,y) - b] = 0 \Leftrightarrow g(x,y) = b \end{cases}$$

$L(x, y, \lambda)$ es **función lagrangiana de** $(\boldsymbol{\rho})$, $\lambda^*_{(\text{Letra griega minúscula lambda})}$ es **Multiplicador de Lagrange asociado a** $(\boldsymbol{x^*, y^*})$. Si (x^*, y^*) es óptimo local de $(\rho) \Rightarrow \exists\, \lambda^*_{(\text{Letra griega minúscula lambda})} \in \mathbb{R}$: (x^*, y^*, λ^*) es un punto crítico de función lagrangiana (punto crítico condicionado).

Los óptimos locales de (ρ) están entre puntos críticos de función lagrangiana:

$$\begin{cases} \left(\dfrac{\partial L}{\partial x}(x, y, \lambda)\right) = 0 \quad (1) \\ \left(\dfrac{\partial L}{\partial y}(x, y, \lambda)\right) = 0 \quad (2) \quad \left(III_{(\text{Condición de 1er orden})}\right) \\ [g(x,y) - b] = 0 \quad (3) \quad \Leftrightarrow \dfrac{\partial L}{\partial \lambda}(x, y, \lambda) = 0 \end{cases}$$

No todas las soluciones $\left(III_{(\text{Condición de 1er orden})}\right)$ serán óptimas, pues requiere condiciones de 2do oren para determinarlos.

Interpretación del Multiplicador de Lagrange: Si (x^*, y^*) es óptimo local de (ρ) con multiplicador asociado $\lambda^*_{(\text{Letra griega minúscula lambda})} \in \mathbb{R}$:

$$\left[\frac{\Delta_{(\text{Letra griega mayúscula delta})} f(x^*, y^*)}{\Delta_{(\text{Letra griega mayúscula delta})}\, b}\right] \approx \lambda^*_{(\text{Letra griega minúscula lambda})}$$

$$\Leftrightarrow \Delta_{(\text{Letra griega mayúscula delta})} f(x^*, y^*) \approx \Delta_{(\text{Letra griega mayúscula delta})}\, b$$

Es decir, **variación de valor óptimo de f se aproxima por** $\boldsymbol{\lambda^*_{(\text{Letra griega minúscula lambda})}}$ **multiplicado por variación de b,** $\boldsymbol{\Delta_{(\text{Letra griega mayúscula delta})}}$.

Para $\boldsymbol{\Delta_{(\text{Letra griega mayúscula delta})}\mathbf{b} = \pm\mathbf{1} \Rightarrow \Delta_{(\text{Letra griega mayúscula delta})}\mathbf{f}(\mathbf{x}^*, \mathbf{y}^*) \approx \pm\,\boldsymbol{\lambda}^*_{(\text{Letra griega minúscula lambda})} \Rightarrow \boldsymbol{\lambda}^*_{(\text{Letra griega minúscula lambda})}}$ **es una aproximación de variación de valor óptimo de f dada una variación unitaria del término independiente b. Definición de segundas Derivadas.**

Las derivadas parciales de segundo orden de $f(x, y)$ son:

- $\partial_{(xx)}f(x,y) = \partial_{(x)}\left[\partial_{(x)}f(x,y)\right] = f_{(xx)}(x,y)$
- $\partial_{(xy)}f(x,y) = \partial_{(x)}\left[\partial_{(y)}f(x,y)\right] = f_{(yx)}(x,y)$
- $\partial_{(yx)}f(x,y) = \partial_{(y)}\left[\partial_{(x)}f(x,y)\right] = f_{(xy)}(x,y)$
- $\partial_{(yy)}f(x,y) = \partial_{(y)}\left[\partial_{(y)}f(x,y)\right] = f_{(yy)}(x,y)$

$$\text{Hess } f(x,y)[H_f(x,y)] = \begin{pmatrix} \partial_{(xx)}f(x,y) & \partial_{(yx)}f(x,y) \\ \partial_{(xy)}f(x,y) & \partial_{(yy)}f(x,y) \end{pmatrix};\ \text{Hess } f(x,y,z)[H_f(x,y,z)]$$

$$= \begin{pmatrix} \partial_{(xx)} & \partial_{(xy)} & \partial_{(xz)} \\ \partial_{(yx)} & \partial_{(yy)} & \partial_{(yz)} \\ \partial_{(zx)} & \partial_{(zy)} & \partial_{(zz)} \end{pmatrix};\ \text{Hess } f(x,y,z,\ldots n)[H_f(x,y,z,\ldots n)]$$

$$= \begin{pmatrix} \partial_{(xx)} & \partial_{(xy)} & \partial_{(xz)} & \cdots & \partial_{(xn)} \\ \partial_{(yx)} & \partial_{(yy)} & \partial_{(yz)} & \cdots & \partial_{(yn)} \\ \partial_{(zx)} & \partial_{(zy)} & \partial_{(zz)} & \cdots & \partial_{(zn)} \end{pmatrix}$$

El determinante Hessiano de $f(x, y)$ se define como:

$$D(x,y) = \left[\left(\partial_{(xx)}f(x,y)\right)\left(\partial_{(yy)}f(x,y)\right)\right] - \left[\left(\partial_{(xy)}f(x,y)\right)\left(\partial_{(yx)}f(x,y)\right)\right]$$

Si derivadas parciales mixtas $\partial_{(yx)}f(x,y), \partial_{(yx)}f(x,y)$ existen y son continuas $\mapsto$ $\partial_{(xy)}f(x,y) = \partial_{(yx)}f(x,y)$ tal que $(:\mapsto)D(x,y) = \left[\left(\partial_{(xx)}f(x,y)\right)\left(\partial_{(yy)}f(x,y)\right)\right] - \left[\left(\partial_{(xy)}f(x,y)\right)^2\right]$.

Teorema (Clasificación de puntos críticos). Suponga que todas las derivadas de 1er orden y 2do orden de $f(x, y)$ existe y (x_0, y_0) es un punto crítico de $f(x, y)$:

- Si $\mathbf{D(x_0, y_0) < 0} \Rightarrow (x_0, y_0)$ es un **punto de silla**.

- Si $\mathbf{D(x_0, y_0) > 0 \Rightarrow (x_0, y_0), \partial_{(xx)} f(x_0, y_0) > 0, f_{(xx)} f(x_0, y_0)}$ **y** $\mathbf{f_{(yy)} f(x_0, y_0)}$ **son** $+$ es un **mínimo relativo**.

- Si $\mathbf{D(x_0, y_0) < 0 \Rightarrow (x_0, y_0), \partial_{(xx)} f(x_0, y_0) < 0, f_{(xx)} f(x_0, y_0)}$ **y** $\mathbf{f_{(yy)} f(x_0, y_0)}$ **son** $-$ es un **máximo relativo**.

- Si $\mathbf{D(x_0, y_0) = 0}$ la información es **no concluyente y requiere otras técnicas de trabajo**.

Criterios de lectura de Hessiana:

a) Si $\forall\, D(x, y, z, \dots, n)$ son$+: \mapsto$ función tiene un **mínimo local en punto crítico.**

Si $\forall\, D(x, y, z, \dots, n)$ son$\pm_{(\text{Alterno e iniciando con valor}-)}: \mapsto$ función tiene un **máximo local en punto crítico.**

Igualando a cero el sistema de ecuaciones vectoriales:

$$\begin{cases} X^t X\hat{\beta}^0 = X^t Y - C^t\lambda \Rightarrow \hat{\beta}^0 = \hat{\beta} - C^t(X^t X)^{-1}\lambda \\ C\hat{\beta}^0 = \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}} \Rightarrow \lambda = S^{-1}\left(\hat{\psi} - \psi_0\right) \end{cases}$$

Si se reemplaza en ecuación anterior, se obtiene resultado buscado. **Lema**. Si C es rango completo:

$$\begin{aligned} &SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)} \\ &\quad = SRC_{(\text{Suma de Residuos al Cuadrado})} \\ &\quad + \left(\hat{\psi} - \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}\right)^t S^{-1} \left(\hat{\psi} - \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}\right) \end{aligned}$$

Tal que $\hat{U}^0 = Y - X\hat{\beta}^0, SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)} = \left[\left(\hat{U}^0\right)^t\left(\hat{U}^0\right)\right]$. **Teorema**. Con hipótesis usuales de método mínimos cuadrados ordinarios (MCO), dado que $C_{(\text{Matriz } q*k \text{ de rango } q < k)}, rg_{(\text{Rango})}(X) = k$, la prueba es de nivel de significancia α y es prueba de razón de máximos de verosimilitud: no se acepta $H_0: C\beta = \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}$ y no se rechaza $H_a: C\beta \neq \psi_{0_{(\text{Vector fijo de } \mathbb{R}^q)}}$ si $F_{(\psi)} =$

$$\left[\frac{(\hat{\psi}-\psi)^t S^{-1}(\hat{\psi}-\psi)}{(q)(S_R^2)}\right] >$$

$F_{(q;\, n-k;\, \alpha)}$:

Fractil de orden $1-\alpha$ de ley de Fisher con $q_{(\text{Grados de libertad del numerador})}, n-k_{(\text{Grados de libertad del denominador})}$

Demostración. Estime razón de máximos de verosimilitud:

$$\lambda(Y) = \left[\frac{\max\limits_{H_0,\, \sigma^2} L_y(\beta, \sigma^2; y)}{\max\limits_{\beta \in \mathbb{R}^k,\, \sigma^2} L_y(\beta, \sigma^2; y)}\right]$$

H_0 no se acepta a nivel de significancia α si y solo si $Pr[\lambda(Y) < c] = \alpha$. Por $L_y(\hat{\beta}, \hat{\sigma}^2) = \left[\left(\frac{1}{(2\pi\hat{\sigma}^2)^{\left(\frac{n}{2}\right)}}\right)\left(e^{\left(-\frac{n}{2}\right)}\right)\right]$:

$$\max_{\beta \in \mathbb{R}^k,\, \sigma^2} L_y(\beta, \sigma^2; y) = \left[\left(\frac{1}{\left(2\pi SRC_{(\text{Suma de Residuos al Cuadrado})}\right)^{\left(\frac{n}{2}\right)}}\right)\left(e^{\left(-\frac{n}{2}\right)}\right)\right]$$

$$\max_{H_0,\, \sigma^2} L_y(\beta, \sigma^2; y) = \left[\left(\frac{1}{\left(2\pi SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)}\right)^{\left(\frac{n}{2}\right)}}\right)\left(e^{\left(-\frac{n}{2}\right)}\right)\right]$$

Tal que $\lambda(Y) = \left(\frac{SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)}}{SRC_{(\text{Suma de Residuos al Cuadrado})}}\right)^{\left(-\frac{n}{2}\right)} \Rightarrow \lambda(Y) < c \Leftrightarrow \left(\frac{SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)} - SRC_{(\text{Suma de Residuos al Cuadrado})}}{SRC_{(\text{Suma de Residuos al Cuadrado})}}\right) > k_{(\text{Constante})}$ donde $k_{(\text{Constante})}$

Por **Lema.** Si C es rango completo, el estimador de mínimos cuadrados bajo hipótesis $H_0: C\beta = \psi_{0\,(\text{Vector fijo de } \mathbb{R}^q)}$ está dado por $\hat{\beta}^0 = \hat{\beta}_{(\text{estimador de mínimos cuadrados sin restricción})} - C^t(X^tX)^{-1}S^{-1}(\hat{\psi} - \psi_0)$ tal que $\hat{\psi} = C\hat{\beta}, S = CC^t(X^tX)^{-1}$ tal que:

$$\left(\frac{\text{SRC}_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)} - \text{SRC}_{(\text{Suma de Residuos al Cuadrado})}}{\text{SRC}_{(\text{Suma de Residuos al Cuadrado})}}\right)$$
$$= \left[\frac{\left(\widehat{\psi} - \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)^t S^{-1}\left(\widehat{\psi} - \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)}{\text{SRC}_{(\text{Suma de Residuos al Cuadrado})}}\right]$$

Tal que:

$$\Pr[\lambda(Y) < c] = \Pr\left[\frac{\left(\widehat{\psi} - \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)^t S^{-1}\left(\widehat{\psi} - \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)}{\text{SRC}_{(\text{Suma de Residuos al Cuadrado})}} > k\right]$$
$$= \Pr\left[\frac{\left(\widehat{\psi} - \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)^t S^{-1}\left(\widehat{\psi} - \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)}{(q)(S_R^2)} > k'\right] = \alpha$$

Por **Corolario**. Bajo hipótesis N, si β_1 está en modelo o si datos han sido centrados $\left(\frac{1}{\sigma^2}\text{SRC}_{(\text{Suma de Residuos al Cuadrado})}\right) \to \chi^2_{(n-2)}$ tal que si β_1 no está en modelo y datos no están centrados, cambian únicamente grados de libertad, que son $n-1$ tal que $k' = F_{\substack{(q;\, n-k;\, \alpha):\\ \text{Fractil de orden } 1-\alpha \text{ de ley de Fisher con } q_{(\text{Grados de libertad del numerador})}, n-k_{(\text{Grados de libertad del denominador})}}}$

3.2.2 A partir de modelo reparametrizado.

Probar hipótesis $H_0: C\beta = \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}$ vs $H_0: C\beta \neq \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}$ es igual a comparar modelo original $Y = X\beta + U$ vs modelo alterno $Y = X\beta^0 + U$ donde β^0 satisface igualdad $C\beta = \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}$ tal que el problema es incluir H_0 en modelo; por lo que, se puede hacer en dos maneras:

1. **Mínimos cuadrados restringidos**. Se minimiza $\|Y - Xb\|^2$ bajo restricción $Cb = \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}$

Por **Lema**. Si C es rango completo, el estimador de mínimos cuadrados bajo hipótesis $H_0: C\beta = \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}$ está dado por $\hat{\beta}^0 = \hat{\beta}_{(\text{estimador de mínimos cuadrados sin restricción})} - C^t(X^tX)^{-1}S^{-1}\left(\widehat{\psi} - \psi_0\right)$ tal que $\widehat{\psi} = C\hat{\beta}, S = CC^t(X^tX)^{-1}$.

2. **Reparametrización de modelo**. Se cambia modelo original por nuevo modelo en que está involucrada H_0; por ejemplo, si $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + u_i$:

a) Si $H_0 = \beta_4 = \beta_5 = 0$ su modelo reparametrizado es $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ donde $q_{(\text{Número de restricciones linealmente independientes})} = 2$ tal que matriz $C_{(\text{Matriz q*k de rango q < k})} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \psi_{0(\text{Vector fijo de } \mathbb{R}^q)} = 0.$

b) Si $H_0 = \begin{cases} \beta_2 = \beta_4 \\ \beta_3 = 0 \end{cases}$ su modelo reparametrizado es $Y_i = \beta_1 + \beta_2(x_{i2} + x_{i4}) + \beta_5 x_{i5} + \varepsilon_i$ tal que $q_{(\text{Número de restricciones linealmente independientes})} = 2$ y matriz $C_{(\text{Matriz q*k de rango q < k})} = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \psi_{0(\text{Vector fijo de } \mathbb{R}^q)} = 0.$

c) Si $H_0: \beta_3 = \beta_5 = 1$ tal que su modelo reparametrizado es $Y_i - x_{i5} = \beta_1 + \beta_2 x_{i2} + \beta_3(x_{i3} - x_{i5}) + \beta_4 x_{i4} + \varepsilon_i$ tal que $q_{(\text{Número de restricciones linealmente independientes})} = 1$ y matriz $C_{(\text{Matriz q*k de rango q < k})} = [0 \quad 0 \quad 1 \quad 0 \quad 1] \psi_{0(\text{Vector fijo de } \mathbb{R}^q)} = 1.$

El **Lema**. Si C es rango completo $SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)} = SRC_{(\text{Suma de Residuos al Cuadrado})} + \left(\widehat{\psi} - \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)^t S^{-1} \left(\widehat{\psi} - \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)$

Tal que $\widehat{U}^0 = Y - X\widehat{\beta}^0, SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)} = \left[\left(\widehat{U}^0\right)^t \left(\widehat{U}^0\right)\right]$ muestra razón F se estima mediante $F_{\left(\psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)} = \left(\frac{SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)} - SRC_{(\text{Suma de Residuos al Cuadrado})}}{(q)(S_R^2)}\right).$

Procedimiento de prueba. Para comprobar $H_0: C\beta = \psi_{0(\text{Vector fijo de } \mathbb{R}^q)}$:

- Ajustar modelo original para obtener $SRC_{(\text{Suma de Residuos al Cuadrado})}$.
- Reparametrizar modelo y ajustarlo para obtener $SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)}$.
- Estimar $F_{\left(\psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)} = \left(\frac{SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)} - SRC_{(\text{Suma de Residuos al Cuadrado})}}{(q)(S_R^2)}\right).$

- Comparar con fractil correspondiente o estimar valor ρ o ρ Value tal que $\hat{\alpha} = \Pr\left[F_{(q;\, n-k):\text{Variable aleatoria Fisher con (n,n−k) grados de libertad}} > F_{\left(\psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)}\right]$.

En ocasiones $SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)}$ puede ser no clara; específicamente si $\psi_{0(\text{Vector fijo de } \mathbb{R}^q)} = 0$ se indica de forma explícita variables incluidas en modelo; por ejemplo $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + u_i$ requiere contrastar $H_0: \beta_3 = \beta_5 = 0$ vs $H_a: \beta_3 \neq 0$ o vs $H_a: \beta_5 \neq 0 \Rightarrow SRC_{(\text{Suma de Residuos al Cuadrado})}(x_2, x_4)$ y $SRC_{(\text{Suma de Residuos al Cuadrado})}(x_2, x_4, x_5)$ en vez de $SRC_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)}$ y $SRC_{(\text{Suma de Residuos al Cuadrado})}$. Además, se puede escribir parámetros en vez de variables:

$$SRC_{(\text{Suma de Residuos al Cuadrado})}(x_2, x_4) = SRC_{(\text{Suma de Residuos al Cuadrado})}(\beta_1, \beta_2, \beta_4)$$

$$SRC_{(\text{Suma de Residuos al Cuadrado})}(x_2, x_3, x_4, x_5) = SRC_{(\text{Suma de Residuos al Cuadrado})}(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$$

La diferencia de sumas de cuadrados, contribución (x_3, x_5) dado que (x_2, x_4) está en modelo se interpreta:

$$SRC_{(\text{Suma de Residuos al Cuadrado})}(x_3, x_5 | x_2, x_4) = SRC_{(\text{Suma de Residuos al Cuadrado})}(x_2, x_4) - SRC_{(\text{Suma de Residuos al Cuadrado})}(x_2, x_3, x_4, x_5)$$

3.3 Prueba de significancia de regresión

Si en regresión $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_k x_{ik} + u_i$ tal que coeficientes $\beta_2, \beta_3, \beta_4, \beta_5 \ldots, \beta_k$ son $\forall k = 0$ o nulos; entonces, Y no depende linealmente de regresores $\{X_j\}$ tal que es importante probar $H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$ vs $H_a: \exists j, j \geq 2, \beta_j \neq 0$ donde H_a o H_1: Y depende linealmente ≥ 1 regresor. Su modelo reparametrizado es $y_i = \beta_1 + \varepsilon_i$ donde el estimador método mínimos cuadrados ordinarios (MCO) de β_1 es $\hat{\beta}_1 = \bar{y}, \hat{y}_i = \bar{y}$ tal que:

$$\mathrm{SRC}_{(\text{Suma de Residuos al Cuadrado})}(\beta_1) = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathrm{SCT}_{(\text{Suma de Cuadrado del Total})}$$

Tal que $q = k - 1$ y razón F sería:

$$F = \left(\frac{\mathrm{SCT}_{(\text{Suma de Cuadrado del Total})} - \mathrm{SRC}_{(\text{Suma de Residuos al Cuadrado})}}{(k-1)(S_R^2)} \right)$$

$$= \left(\frac{\frac{\mathrm{SEC}_{(\text{Suma de Estimaciones al Cuadrado})}}{(k-1)}}{(S_R^2)} \right)$$

Definición. Se definen medias o promedios de sumas de cuadrados:

$$\mathrm{MSEC}_{(\text{Media de Suma de Estimaciones al Cuadrado})} = \left(\frac{\mathrm{SEC}_{(\text{Suma de Estimaciones al Cuadrado})}}{(k-1)} \right)$$

$$\mathrm{MSRC}_{(\text{Media de Suma de Residuos al Cuadrado})} = \left(\frac{\mathrm{SEC}_{(\text{Suma de Estimaciones al Cuadrado})}}{(n-k)} \right) = S_R^2$$

Estos resultados se escriben en tabla ANOVA, ADEVA, ANDEVA, ANVA, AVAR o ANOVA de Fisher:

Tabla 23. *ANOVA de Fisher*

Análisis de Varianza o **ANOVA**, **ADEVA**, **ANDEVA**, **ANVA**, **AVAR o ANOVA de Fisher**					
F. de V$_{\text{(Factor de Variación)}}$ o **VF**$_{\text{(Variation factor)}}$	**Gl**$_{\text{(Grados de Libertad)}}$ o **Df**$_{\text{(Degrees of freedom)}}$	**SC**$_{\text{(Suma de Cuadrados)}}$ o **Sum Sq**$_{\text{(Sum of square)}}$	**CM**$_{\text{(Cuadrado Medio)}}$ o **Mean Sq**$_{\text{(Mean squares)}}$	**F**$_{\text{(calculada)}}$ o **F value**$_{\text{(Calculated F)}}$	**Pr**$(>\mathbf{F})$ o $\boldsymbol{\rho - \text{Value}}$
Total	$(n-1)$	$STC_{\text{(Suma Total de Cuadrados)}} = \|Y - 1_n\bar{Y}\|^2$			
Intergrupo o Regresión	$(k-1)_{\text{(gl numerador)}}$	$SEC_{\text{(Suma de Estimaciones al Cuadrado)}} = \left\|\hat{Y} - 1_n\bar{Y}\right\|^2$	$MSEC/gl_{\text{(numerador)}}$	$\frac{MSEC_{\text{Regresión}}}{MSRC_{\text{Error}}}$	$\frac{gl_{\text{(numerador}}}{gl_{\text{(denominado}}}$
Intragrupo o Error	$(n-k)_{\text{(gl denominador}}$	$SRC_{\text{(Suma de Residuos al Cuadrado)}} = \left\|Y - \hat{Y}\right\|^2$	$MSRC/gl_{\text{(denominador)}}$		

La sucesión de pruebas $\beta_j = 0, j = 2,3,4,5, \dots, k \neq$ prueba simultánea $\beta_2 = \beta_3 = \cdots = \beta_k = 0$. Por ejemplo, es posible que todas las pruebas $\beta_j = 0$ sean no significativas de forma individual, con nivel de significancia α, pero razón F de prueba simultánea puede ser significativa a nivel α.

3.4 Hipótesis para significación de modelo

Si estadístico F de Snedecor es, asociado a un nivel crítico de probabilidad de obtener valores, mayor al nivel de error asumido, habitualmente 95 % de confiabilidad estadística, error ó nivel de significancia (α) igual a 0.05, se interpreta "con base en resultados obtenidos de datos muestrales (poblacionales), no se acepta hipótesis nula, que indica ninguna variable exógena es buen predictor de Y ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \cdots = \mu_k = \mu\ (0)$ o $H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = \cdots \tau_k = 0$) y no se rechaza la alternativa, que indica al menos una variable exógena es buena predictor de Y respecto al resto ($H_a: \mu_i \neq \mu_j$ para algún $i \neq j$ o $H_a: \tau_i \neq 0$ para algún i)".

Con diferencia parcial respecto a (Gujarati & Porter, 2010) que afirman "con base en una prueba de significancia, como prueba t, se dice "aceptar" la hipótesis nula, todo lo que se afirma es que, con base en la evidencia dada por la muestra, no existe razón para rechazarla, no se sostiene que la hipótesis nula sea verdadera con absoluta certeza" y, de acuerdo con (González B. G., 1985), "de la misma manera que un tribunal se pronuncia un veredicto de "no culpable" en vez de "inocente", así la conclusión de una prueba estadística es **"no rechazar" en vez de "aceptar"**.

Según (Wackerly, Mendenhall, & Scheaffer, 2010), las pruebas de hipótesis se pueden abordar desde una perspectiva de Bayes con base en información previo a datos acerca de un parámetro para su distribución posterior. En consecuencia, se está interesado en probar que el parámetro θ se encuentra en uno de dos conjuntos de valores Ω_0 y Ω_a.

Cuando se pruebe $H_0: \theta \in \Omega_0$ contra $H_a: \theta \in \Omega_a$, usando método de cálculo de probabilidades posteriores sería $P*(\theta \in \Omega_0)$ y $P*(\theta \in \Omega_a)$ aceptando la hipótesis con más alta probabilidad posterior; es decir, **aceptar $\mathbf{H_0}$ si $\mathbf{P*(\theta \in \Omega_0) > P*(\theta \in \Omega_a)}$** o **aceptar $\mathbf{H_a}$ si $\mathbf{P*(\theta \in \Omega_a) > P*(\theta \in \Omega_0)}$**.

3.5 Pruebas respecto a parámetros $\boldsymbol{\beta}$ individuales

En modelo pueden estar incluidas variables que son redundantes o no aportan significativamente a la respuesta, eso se puede comprobar realizando una prueba sobre los parámetros individuales de regresión.

Entonces, se tiene la siguiente prueba de hipótesis para un parámetro $\beta_{(i)}$ fijo.

1. Hipótesis Nula. $H_0: \beta_{(k)} = 0$.
2. Hipótesis Alternativa. $H_1: \beta_{(k)} \neq 0$.
3. Estadístico de Prueba. $t_{(Calculada)} = \left[\frac{b_{(i)}}{s\sqrt{C_{(ii)}}}\right]$.
4. Región de rechazo. No se acepta H_0 si $t_{(Calculada)} \leq t_{\left(n-k-1;\frac{\alpha}{2}\right)}$ o $t_{(Calculada)} \geq t_{\left(n-k-1;\frac{\alpha}{2}\right)}$

3.6 Intervalos de confianza

En la regresión múltiple se pueden formular estimaciones por intervalo para los valores de los coeficientes de regresión, para los valores estimados y para nuevas predicciones.

3.6.1 Para $\boldsymbol{\beta}$

Puesto que el vector de coeficientes βes desconocido, los consideramos como una variable ale Multivalente, normalmente distribuida con media b y matriz de covarianza $\sigma^2(xtx) - 1$, por 1 cada uno de los estadísticos:

$$\frac{bj - \beta j}{s\sqrt{C}jj} \; j = 0,1,\ldots,k$$

Sigue una ley t con $(n-k-1)$ grados de libertad y donde C_{jj} es el j—esimo elemento de la matriz $(xt\,x)-1$. Un intervalo de confianza al $100(1-\alpha)$ % para el coeficiente de regresión $\beta = 0,1,\ldots,k)$, es $(b_0 - t\alpha/2\,(n-k\,1)\,s\sqrt{C_{jj}};\ b_j + t\,\alpha/2\,(n-k-1)\,s\sqrt{C_{jj}})$.

3.6.2 Estimación (E[Y])

Si se desea conocer el intervalo de confianza para la respuesta media de un punto en particular $x,\ x_{p2,\ldots,}x_{pk,}$ definimos el vector

$$X_P = \begin{pmatrix} 1 \\ x_{p1} \\ x_{p2} \\ . \\ . \\ . \\ x_{pk} \end{pmatrix}$$

La respuesta en este punto es estimador es insesgado; es decir, $E(\bar{Y}p) = xtp\,\beta)$ y su varianza es $Var(\bar{Y}p) = ò2xtp(xtx) - 1xp.$

Un intervalo de confianza al $100(1-\alpha)\%$ para la estimulación, E (yp). en el punto x_p es:

$$(\bar{Y}p - t\alpha/2(n-k-1)s\sqrt{x}tp(xtx) - 1xp;\ \bar{Y}p + t\alpha/2(n-k-1)s\sqrt{x}tp(xtx) - 1xp)$$

3.6.3 Predicción (Y_P)

Un modelo de regresión se aplica en la realización de pronóstico correspondientes a valores particulares de las variables independientes, x_p. La respuesta en este punto es $\bar{Y}p = xptb$. El intervalo de confianza de nivel $100(1-\alpha)\%$ para la predicción, y_p, es:

$$(\bar{Y}p - t\alpha/2(n-k-1)s\sqrt{x}tp(xtx) - 1xp;\ \bar{Y}p + t\alpha/2(n-k-1)s\sqrt{x}tp(xtx) - 1xp)$$

3.7 Coeficientes de regresión lineal múltiple

3.7.1 Interpretación

La ecuación de superficie de regresión, estimación de relación funcional buscada, es:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \cdots + \hat{\beta}_k x_k$$

El coeficiente $\hat{\beta}_k$ de x_k es efecto de variable X_k cuando los demás regresores permanecen constantes tal que si X_k se incrementa en una unidad y regresores $X_2, X_3, X_4, \ldots, X_{j-1}, X_j, \ldots, X_k$ permanecen constantes tal que Y se incrementa o decrementa en promedio $\hat{\beta}_k$ unidades.

Entonces, $\hat{\beta}_k$ se considera efecto parcial o directo de X_k cuando se controla o elimina efectos de otras variables explicativas o exógenas. Para interpretar el coeficiente de una variable es necesario conocer el resto de variable explicativas o exógenas incluidas en el modelo. Si se ejecuta la regresión:

$$Y_i = \alpha_1 + \alpha_j x_{ij} + \varepsilon_j$$

$\hat{\alpha}_j$ es un coeficiente que representa el efecto total de X_k respecto a Y; es decir, es la suma de efectos directo e indirecto de X_k respecto a Y mediante las otras variables.

En general, a excepción de ortogonalidad de regresores $\hat{\alpha}_k \neq \hat{\beta}_k$; aunque, la diferencia se debe a correlaciones entre regresores, $\hat{\beta}_k$ indica efecto de X_k luego de eliminar efecto del resto de regresores se considera efecto directo.

3.7.2 Interpretación geométrica

La media $E(Y) = X\beta$ se estima mediante $\widehat{Y} = X\hat{\beta} = \left[H_{(\text{Matrix hat o "sombrero"})}\right]Y \equiv [X(X^tX)^{-1}X^t]Y$, su diagonal desempeña un rol importante en detección de puntos palanca.

Teorema. Sea $H_{(\text{Matrix hat o "sombrero"})} = X(X^tX)^{-1}X^t$, vector $\widehat{Y} = X\hat{\beta} = \left[H_{(\text{Matrix hat o "sombrero"})}\right]Y$ es proyección ortogonal respecto a producto escalar usual de Y sobre $F = \text{img}_{(\text{Imagen})}(X)$. Vectores $\widehat{Y}, \widehat{U}$ son ortogonales (proyección ortogonal Y sobre F):

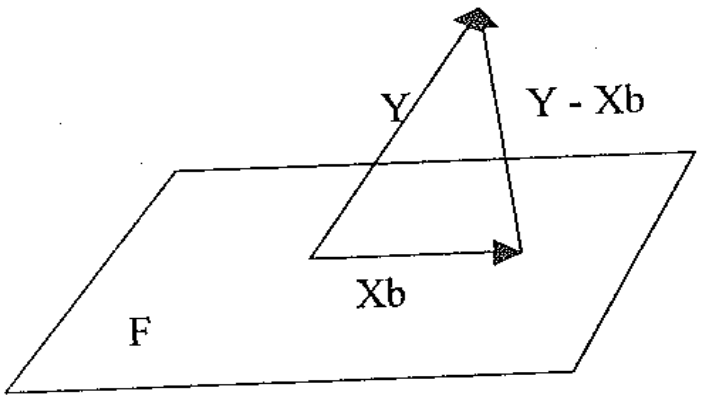

Figura 4. *Proyección ortogonal respecto a producto escalar usual de Y*

Demostración. Sería suficiente demostrar que matriz $H_{(\text{Matrix hat o "sombrero"})}$ es un proyector ortogonal tal que es suficiente que P sea simétrica e idempotente. $H_{(\text{Matrix hat o "sombrero"})}$ es simétrica tal que la igualdad muestra que es idempotente:

$$H^2_{(\text{Matrix hat o "sombrero"})} = [X(X^tX)^{-1}X^t][X(X^tX)^{-1}X^t] = [X(X^tX)^{-1}X^t]$$
$$= H_{(\text{Matrix hat o "sombrero"})}$$

Observación. La norma de vector de residuos $\widehat{U} = Y - \widehat{Y}$ es mínima si y sólo si vectores $\widehat{Y}$ y $\widehat{U}$ son ortogonales. Repaso a Álgebra Lineal:

Vectores $(u_1, u_2, u_3, u_4, \dots, u_k)$ de $\mathbb{R}^n$ son linealmente independientes ssi:

$$a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4 + \cdots + a_ku_k = 0$$

Implica que todos los coeficientes $a_j = 0$ o nulos. Si no, son linealmente independientes. Un conjunto infinito B de vectores es linealmente independiente ssi todo conjunto B de vectores es linealmente independiente ssi todo sub conjunto finito de B es linealmente independiente.

Si vectores $\{u_j\}$ son linealmente dependientes existe, en igualdad anterior, un coeficiente $a_j \neq 0$ o no nulo tal que el vector correspondiente a dicho coeficiente es combinación lineal del resto de vectores. Por ejemplo: $a_j \neq 0 \Rightarrow u_k = \left(-\frac{1}{a_k}\right)(a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4 + \cdots + a_{k-1}u_{k-1})$.

Bases y dimensión. Sea E un espacio vectorial y B un sub conjunto de E:

Si B genera a E ssi ∀ vector de E es una combinación lineal de vectores de B.

Si B es linealmente independiente se dice que B es una base de E.

Si B es una base de E, el número de elementos de B se llama dimensión de B, si B es infinito se dice que E es dimensión infinita.

Número de filas linealmente independientes, en una matriz $A_{(n*m)}$ es igual a número de columnas linealmente independientes.

Sea A una matriz $A_{(n*m)}$. El número de filas o columnas linealmente independientes se llama rango de matriz y se denota como $rg(A_{(Rango)})$.

De bases y dimensión. Sea E un espacio vectorial y B un sub conjunto de E, es obvio que $rg(A_{(Rango)}) \leq \min\{n, m\}$. Si se logra igualdad $[rg(A_{(Rango)}) \leq \min\{n, m\}]$ se dice que matriz A es rango completo.

Matriz cuadrada A es simétrica ssi al cambiar filas por columnas la matriz A no cambia; es decir, ssi $A^t = A$ donde A^t es transpuesta de A.

Matriz cuadrada A es diagonal ssi es cuadrada y todos los elementos fuera de la diagonal principal son 0 o nulos:

$$A = \text{diag}_{(\text{Diagonal})}\{a_1 + a_2 + a_3 + a_4 + \cdots + a_n\} = \begin{bmatrix} a_1 & 0 & 0 & 0 & 0 & & 0 \\ 0 & a_2 & 0 & 0 & 0 & & 0 \\ 0 & 0 & a_3 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & a_4 & 0 & & 0 \\ 0 & 0 & 0 & 0 & a_5 & & 0 \\ & & \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & a_n \end{bmatrix}$$

Matriz identidad es matriz diagonal con todos $a_i = 0$:

$$I = \text{diag}_{(\text{Diagonal})}\{1,1,1,1, \dots ,1\}$$

Cuando requiere indicar el orden de matriz se estribe I_n en vez de I.

Si A matriz cuadrada $(n * n)$ es no singular ssi existe una matriz B cuadrada $(n * n)$:

$$AB = BA = I$$

Matriz B se llama inversa de A y se escribe A^{-1}

$(AB)^t = B^t A^t$. Si matrices (A, B) son no singulares y de igual dimensión $(AB)^{-1} = B^{-1}A^{-1}$.

Una matriz cuadrada $A_{(n*n)}$ es simétrica semi definida positiva, escrita como $A_{(n*n)} \geq 0$, ssi es simétrica y $\forall x, x \in \mathbb{R}^n, x^t Ax \geq 0$ se dice simétrica definida positiva, escrita $A > 0$, ssi es se mi definida positiva y $x^t Ax = 0 \Rightarrow x = 0$.

Sea E, F dos espacios vectoriales sobre un mismo cuerpo de escalares K.

Una función $f: E \longrightarrow F$ se dice aplicación lineal ssi $\forall a \in k, \forall x, y \in E$ se tiene que $f(ax + y) = af(x) + f(y)$.

$\forall A_{(m*n)}$ define una aplicación lineal $\mathbb{R}^n$ en $\mathbb{R}^m$:

$$A: \mathbb{R}^n \rightarrow \mathbb{R}^m: x \rightarrow Ax$$

A su vez toda aplicación lineal de $\mathbb{R}^n$ en $\mathbb{R}^m$ define una matriz $(m * n)$.

Si A es aplicación lineal de $\mathbb{R}^n$ en $\mathbb{R}^m$ se llama espacio imagen de A al conjunto:

$$\mathrm{img}_{(\mathrm{Imagen})}(A) = \{w \in \mathbb{R}^m | w = Au, u \in \mathbb{R}^n\}$$

Se llama núcleo de A al conjunto:

$$\mathrm{Ker}(A) = \{u \in \mathbb{R}^n | Au = 0\}$$

Se puede demostrar que $\dim_{(\mathrm{Dimensión})}\left[\mathrm{img}_{(\mathrm{Imagen})}(A)\right] = \mathrm{rg}\left(A_{(\mathrm{Rango})}\right)$ y $\dim_{(\mathrm{Dimensión})}\left[\mathrm{img}_{(\mathrm{Imagen})}(A)\right] + \dim_{(\mathrm{Dimensión})}\left[\mathrm{Ker}_{(\mathrm{Kernel\ o\ núcleo})}(A)\right] = n$

Si $A_{(n*n)}$ es matriz cuadrada $(n * n)$ simétrica definida positiva, se define el producto escalar de dos vectores u, v de $\mathbb{R}^n$ respecto de A por $\langle u, u\rangle_{(A)} = (u^t)(Au)$ se define norma de v respecto de A por $\|u\|_{(A)} = \sqrt{(u^t)(Au)}$.

Si A es matriz identidad no hace falta sub índices tal que se dice que son producto escalar y norma usuales.

Si A es matriz cuadrada $(n * n)$ se llama traza de A a suma de elementos de su diagonal principal tal que $\mathrm{tr}_{(\mathrm{Traza\ de\ A})}(A) = \sum_{i=1}^{n}(a_{ii})$.

Sea A matriz $(n * m)$, B matriz $(m * n)$ tal que $\mathrm{tr}_{(\mathrm{Traza\ de\ AB})}(AB) = \mathrm{tr}_{(\mathrm{Traza\ de\ BA})}(BA)$.

3.8 Coeficientes de determinación

3.8.1 Determinación

El coeficiente de determinación se emplea como una medida de la educación del modelo e informa sobre fuerza de relación existente entre variables independientes y dependiente.

3.8.2 R^2

Coeficiente de determinación se define:

$$R^2 = \left(\frac{SEC_{(\text{Suma de Estimaciones al Cuadrado})}}{STC_{(\text{Suma Total de Cuadrados})}}\right)$$

En siguiente definición es importante recordar que si regresión tiene término constante, $\sum \hat{u}_i^2 = 0 \equiv (\bar{y} - \hat{y})$.

Definición. Coeficiente de correlación múltiple entre Y i $(X_2, X_3, X_4, X_5 \dots, X_k)$ y coeficiente simple entre Y i $\hat{Y}$:

$$R_{(Y.2,3,4,5,\dots,k)} = \left[\frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_i (y_i - \bar{y})^2 (\hat{y}_i - \bar{y})^2}}\right]$$

Teorema. Si regresión tiene término $R^2 = R_{(Y.2,3,4,5,\dots,k)}$.

Demostración:

$$\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y}) = \sum_i (y_i \hat{y}_i) - (n\bar{y}^2) = \hat{Y}^t Y - n\bar{y}^2 = SEC_{(\text{Suma de Estimaciones al Cuadrado})}$$

Tal que:

$$\begin{aligned} R_{(Y.2,3,4,5,\dots,k)} &= \left[\frac{SEC^2_{(\text{Suma de Estimaciones al Cuadrado})}}{\left(STC_{(\text{Suma Total de Cuadrados})}\right)\left(SEC_{(\text{Suma de Estimaciones al Cuadrado})}\right)}\right] \\ &= \left[\frac{\left(SEC_{(\text{Suma de Estimaciones al Cuadrado})}\right)\left(\cancel{SEC_{(\text{Suma de Estimaciones al Cuadrado})}}\right)}{\left(STC_{(\text{Suma Total de Cuadrados})}\right)\left(\cancel{SEC_{(\text{Suma de Estimaciones al Cuadrado})}}\right)}\right] \\ &= \left[\frac{\left(SEC_{(\text{Suma de Estimaciones al Cuadrado})}\right)}{\left(STC_{(\text{Suma Total de Cuadrados})}\right)}\right] \end{aligned}$$

Teorema. Considérese probar hipótesis $H_0: C\beta = \psi_{0\,(\text{Vector fijo de } \mathbb{R}^q)}$ vs $H_1: C\beta \neq \psi_{0\,(\text{Vector fijo de } \mathbb{R}^q)}$.

Sea R^2 coeficiente de correlación en modelo original, completo y R_0^2 coeficiente de correlación correspondiente a modelo reparametrizado, restringido:

$$\left[\frac{\left(\mathrm{SRC}_{(0:\text{ Suma de Residuos al Cuadrado bajo } H_0)} - \mathrm{SEC}_{(\text{Suma de Estimaciones al Cuadrado})}\right)}{\left(\mathrm{SEC}_{(\text{Suma de Estimaciones al Cuadrado})}\right)}\right] = \left[\frac{(R^2 - R_0^2)}{(1 - R^2)}\right]$$

Tal que:

$$F_{\left(\psi_{0(\text{Vector fijo de } \mathbb{R}^q)}\right)} = \left[\left(\frac{n-k}{q}\right)\left(\frac{(R^2 - R_0^2)}{(1 - R^2)}\right)\right]$$

En caso de ambigüedad se incluir nombre de variables o parámetros en notaciones R^2 y R_0^2.

Observación. Se puede demostrar que $y_i = \beta_1 + \varepsilon_i$, $R_0^2 = 0$ tal que razón F para prueba de significación de regresión se estima mediante:

$$F = \left[\left(\frac{n-k}{k-1}\right)\left(\frac{R^2}{1 - R^2}\right)\right]$$

R^2será significativo si y solo si razón F es significativa a nivel α:

$$\left[\left(\frac{n-k}{k-1}\right)\left(\frac{R^2}{1 - R^2}\right)\right] > F_{(k-1;n-k;\,\alpha)} \Leftrightarrow R^2 > \left[\frac{F_{(k-1;n-k;\,\alpha)}(k-1)}{n - k + F_{(k-1;n-k;\,\alpha)}(k-1)}\right]$$

Corolario. Coeficiente R^2 incrementa con número de regresores en modelo.

Demostración. Razón F siempre es positiva tal que $R^2 > R_0^2$. En caso de un solo regresor $(k = 2)$, el coeficiente de correlación simple, empírico, es igual a coeficiente de correlación múltiple. Este coeficiente es una cantidad estrechamente relacionada con el concepto de Coeficiente de Determinación r^2 (dos variables) ó R^2 (regresión múltiple) como medida de "bondad de ajuste".

El **Coeficiente de Correlación Muestral** explica el grado de asociación, mide la fuerza o grado de asociación lineal, entre dos variables. Su cálculo es a partir de:

$$r = \pm\sqrt{r^2} = \frac{\sum X_i Y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} = \frac{n\sum X_i Y_i - [(\sum X_i)(\sum Y_i)]}{\sqrt{[n\sum x_i^2 - (\sum X_i)^2][n\sum y_i^2 - (\sum Y_i)^2]}}$$

El **Coeficiente de Correlación Poblacional** $\boldsymbol{\rho}_{\text{(letra griega minúscula Rho o ro)}}$ es:

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{[\text{Var}\,(X)\text{Var}\,(Y)]}} = \frac{\text{Cov}(X, Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Entonces, ρ es una medida de Asociación Lineal entre dos variables y su valor se sitúa entre ± 1, donde -1 indica una perfecta asociación negativa y $+1$ es una perfecta asociación positiva.

Con base en lo anterior, se deduce:

$$\text{Cov}(X, Y) = \rho(\sigma_X \sigma_Y)$$

Algunas propiedades de r son:

I. Tener signo positivo o negativo, según el signo del término en numerador de $r = \pm\sqrt{r^2} = \frac{\sum X_i Y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} = \frac{n\sum X_i Y_i - [(\sum X_i)(\sum Y_i)]}{\sqrt{[n\sum x_i^2 - (\sum X_i)^2][n\sum y_i^2 - (\sum Y_i)^2]}}$ que mide la Covariación Muestral de dos variables.

II. Se ubica entre límites -1 y $+1$; es decir, $-1 \leq r \leq 1$.

III. Es simétrico por naturaleza; es decir, el Coeficiente de Correlación entre X y Y, denotado por r_{XY}.

IV. Es independiente del origen y escala; es decir, si se define $X_i^* = aX_i + C$ y $Y_i^* = bY_i + d$ tal que $a > 0$, $b > 0$, c y d son constantes. Entonces, r entre X^* y Y^* es igual a r entre variables originales X y Y.

V. Si X y Y son estadísticamente independientes, pues dos variables aleatorias X y Y son estadísticamente independientes si y sólo si ($\Leftrightarrow$) $f(x, y) = f(x)\, f(y)$; es decir, si Función de Densidad de Probabilidad conjunta se expresa como el producto de las FDP marginales, el Coeficiente de Correlación ente ellas es cero, aunque no significa que dos

variables sean independientes. En otras palabras, una correlación igual a cero no necesariamente implica independencia.

VI. Es una medida de Asociación Lineal o Dependencia Lineal solamente, su uso es la descripción de relaciones no lineales no tiene significado. Así, en la siguiente figura h, $Y = X^2$ es una relación exacta y $r = 0$, pues es una expresión cuadrática (parábola positiva) y, en consecuencia, una relación no lineal.

VII. Es una medida de asociación lineal entre dos variables y no implica obligatoriamente una relación causa-efecto: "Una relación estadística por más fuerte y sugerente que sea nunca podrá establecer una conexión causal, por lo que ideas de causalidad provendrán de estadísticas externas y, finalmente, de una u otra teoría". Por lo tanto, "una relación estadística por sí misma no puede, lógicamente, implicar causalidad".

En el contexto de regresión, r^2 es una medida con más significado que r, pues la primera indica la proporción de variación en variable dependiente, explicada, predicha, regresada, respuesta, endógena, resultado o controlada explicada por la (s) variable (s) independiente (s), explicativa (s), predictora (s), regresora (s), estímulo (s), exógena (s), covariante (s) o de control (s).

En consecuencia, constituye una medida global del grado en que la variación en una variable determina la variación de otra. r no tiene este valor y, además, la interpretación de r (R) en un modelo de regresión múltiple es de valor dudoso.

3.8.3 Corregido $\bar{R}^2$

Si coeficiente R^2 incrementa con número de regresores en modelo es conveniente penalizarlo.

Definición. Coeficiente de correlación es definido como:

$$\overline{R}^2 = 1 - \left[\left(\frac{n-1}{n-k}\right)(1-R^2)\right]$$

Coeficiente $\overline{R}^2$ se obtiene:

$$R^2 = \left(\frac{SEC_{(\text{Suma de Estimaciones al Cuadrado})}}{STC_{(\text{Suma Total de Cuadrados})}}\right)$$

$$1 - R^2 = \left(\frac{SRC_{(\text{Suma de Residuos al Cuadrado})}}{STC_{(\text{Suma Total de Cuadrados})}}\right)$$

$$1 - \overline{R}^2 = \left[\frac{\left(\frac{SRC_{(\text{Suma de Residuos al Cuadrado})}}{n-k}\right)}{\left(\frac{STC_{(\text{Suma Total de Cuadrados})}}{n-1}\right)}\right]$$

Teorema. $\overline{R}^2 < R^2$.

3.9 Parcial

Dadas dos variables (X, Y) el coeficiente de correlación simple es na medida de relación lineal entre ellas, pero la pregunta es si medirá de forma efectiva su relación lineal cuando existen otras variables que están asociadas con estas; aunque, la respuesta es no. Por ejemplo: considere variables velocidad, potencia y peso de un vehículo. Existen vehículos pequeños demasiado veloces, pero también existen grandes, como buses o camiones que igual son veloces.

Estas observaciones indican una correlación no significativa entre velocidad y peso tal que contradice el sentido común. En realidad, sucede que vehículos más pesados también son más potentes, no existen camiones con potencia de un auto normal. El cálculo verdadero de correlación entre velocidad y peso se debería mantener constante la potencia.

El coeficiente de correlación parcial mide la correlación entre dos variables, manteniendo las demás con valores constantes. Si un conjunto de variables es

$\{X_2, X_3, X_4, X_5, \dots, X_k\}$, el coeficiente de correlación parcial entre dos de ellas es una medida, adimensional, de su relación lineal luego de eliminar el efecto del resto de variables de ambas. Se conoce que $Y = \widehat{Y} + \widehat{U}$: $\widehat{Y}$ es parte de Y explicada por regresores mientras que el residuo $\widehat{U}$ es parte no explicada de Y por regresores tal que el coeficiente de correlación parcial se definirá mediante residuos.

Definición. Por cuestiones de facilidad el coeficiente de correlación parcial entre (X_2, X_3) dado el resto de variables. Sea $\widehat{U}_{(2.34\dots k)}$ el vector de residuos de regresión de X_2 en $\{X_4, X_5, \dots, X_k\}$, $\widehat{U}_{(3.45\dots k)}$ vector de residuos de regresión de X_3 en $\{X_4, X_5, \dots, X_k\}$.

El coeficiente de correlación parcial entre (X_2, X_3) dado por $\{X_4, X_5, \dots, X_k\}$ es coeficiente de correlación simple entre residuos de dos regresiones:

$$r_{23.45,\dots,k} = \left[\frac{\sum_i \hat{u}_{2.45,\dots,k(i)} \hat{u}_{3.45,\dots,k(i)}}{\sqrt{\sum_i \hat{u}^2_{2.45,\dots,k(i)} \sum_i \hat{u}^2_{3.45,\dots,k(i)}}} \right]$$

Número de variables eliminadas, fijas, $\{X_4, X_5, \dots, X_k\}$ es orden del coeficiente de correlación parcial. r_{23} es orden cero, $r_{23.4}$ es orden 1, $r_{23.34}$ es orden 2, etcétera. Se define correlación parcial entre (X_2, X_3), pero igual se define para $\forall(X_i, X_j)$ incluida variable dependiente.

El coeficiente de correlación parcial tiene sentido sólo cuando se precisas todas las variables tal que se requiere todos los sub índices.

Teorema. Coeficiente de correlación parcial entre variable dependiente Y y un regresor X_j $(r_{yj.R})$ se puede estimar con base en razón t de Student:

$$r^2_{(yj.R)} = \left[\frac{t_j^2}{t_j^2 + (n-k)} \right]$$

Teorema. Coeficientes de correlación parcial de orden p se pueden estimar en forma recursiva con base en coeficientes de correlación de orden cero:

$$r_{(12.3)} = \left[\frac{r_{(12)} - r_{(13)}r_{(23)}}{\sqrt{\left(1 - r^2_{(13)}\right)\left(1 - r^2_{(23)}\right)}}\right]$$

$$r_{(12.34,\ldots,p)} = \left[\frac{r_{(12.34,\ldots,(p-1))} - r_{(1p.34,\ldots,(p-1))}r_{(2p.34,\ldots,(p-1))}}{\sqrt{\left(1 - r^2_{(1p.34,\ldots,(p-1))}\right)\left(1 - r^2_{(2p.34,\ldots,(p-1))}\right)}}\right]$$

Observación. La fórmula $r_{(12.3)} = \left[\frac{r_{(12)} - r_{(13)}r_{(23)}}{\sqrt{\left(1 - r^2_{(13)}\right)\left(1 - r^2_{(23)}\right)}}\right]$ evidencia que correlación parcial puede ser diferente de cero incluso cuando coeficiente de correlación simple es nulo.

5.1.1. Múltiple

El coeficiente de determinación múltiple se define:

$$R^2 = \frac{SRC_{(\text{Suma de Residuos al Cuadrado})}}{SC_{(\text{Suma Cuadrados})(vy)}} = 1 - \frac{SRC_{(\text{Suma de Residuos al Cuadrado})}}{SC_{(\text{Suma Cuadrados})(vy)}}$$

El rango de la variable de R^2 es $0 \leq R^2 \leq 1$. Un valor alto de R^2, cercano a 1, indica que modelo de regresión es bueno y si R^2 es cercano a 0, el ajuste es malo. Si $R^2 = 0$ señala falta por completo el ajuste del modelo a los datos y si $R^2 = 1$ se tiene un ajuste perfecto.

Puesto que R^2 tiende a sobreestimar el valor de correlación entre variables involucradas, se emplea el coeficiente de determinación ajustado, R^2, que se et diseñado para compensar el sesgo optimista de R^2. El coeficiente de determinación ajustado se define:

$$R_{\alpha}^{2} \ = R^{2} - \frac{k(1 - R^{2})}{n - k - 1}$$

Equivalente, R_{α}^{2} se calcula:

$$R_{\alpha}^{2} = 1 - \frac{SCE_{(\text{Suma de Cuadrados del Error})}/(n - k - 1)}{SC_{(\text{Suma Cuadrados})(yy)}/(n - 1)},$$

El rango de variación de R_{α}^{2} es $0 \leq R_{\alpha}^{2} \leq 1$ y su interpretación es la misma que la del coeficiente determinación múltiple R^{2}.

CAPÍTULO IV

CAPÍTULO IV

REGRESIÓN POLINOMIAL

El primer paso para escoger un modelo que describa los datos es la realización de una dispersión de observaciones. La relación sugerida por datos es la que permite escoge el modelo que los describa adecuadamente. Cuando los datos presentan un esquema de comportamiento curvilíneo puede ser necesario un modelo de tipo polinomial para datos:

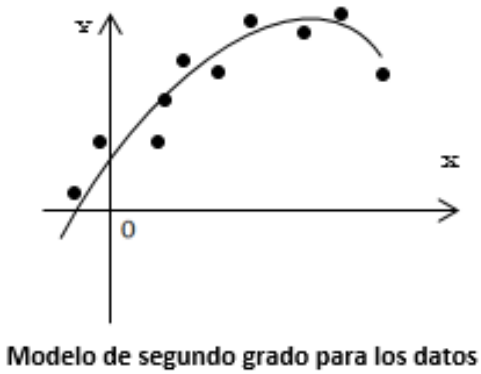

Figura 5. Realización de una dispersión de observaciones

Para transformar un modelo polinomial en uno de regresión multiplique se escoge un modelo de segundo grado en una variable:

$$y_i = \beta_1 + \beta_2 x_1 + \beta_2 x_2 + e_i$$

Si se hace situaciones de variables $x_1 = x^1$ y $x_2 = x^2$ la ecuación sería:

$$y_i = \beta_1 + \beta_2 x^1 + \beta_2 x^2 + e_i$$

Es un modelo de regresión múltiple en dos variables. Se está en posibilidad de aplicar la teoría descrita. En general, si se tiene un modelo polinomial con una variable explicativa

$$y_i = \beta_1 + \beta_2 x^1 + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \cdots + \beta_k x^k + e_i$$

Se puede convertir en uno de regresión múltiple mediante la transformación $x_i + x^i$:

$$y_i = \beta_1 + \beta_2 x_1 + \beta_2 x^2 + \beta_3 x_3 + \beta_4 x^4 + \cdots + \beta_k x_k + e_i$$

Otros modelos polinomiales que incluyen más de un variable, que pueden transformarse a una línea múltiple, son los polinomios en variables, como el de segundo grado en dos variables:

$$y_i = \beta_1 + \beta_2 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + e_i$$

Cuando se ajusta un modelo polinomial es preciso escoger el polinomio de menor grado posible tal que se deberán realizar reiteradas pruebas de hipótesis, en las que se fijaran aquellas variables que se han de incluir y excluir en modelo final.

4.1 Regresión con variables cualitativas

En los modelos tratados se empleó variables independientes de naturaleza cuantitativa; es decir, se expresan numéricamente y son el resultado de mediciones instrumentales. Si se desea incorporar en modelo una variable cualitativa será necesario introducir variables indicadoras, ficticias, que permiten diferenciar los distintos niveles que toma tal variable; por ejemplo, una variable X que indiquen la estación del año puede ser definida:

$$X = \begin{cases} 0: \text{Es invierno} \\ 1: \text{Es verano} \end{cases}$$

En general, una variable cualitativa con t niveles se representa mediante $t - 1$ variables indicadoras, se asignan valores de 0 y 1 llamadas dicotómicas, binomiales o dummy.

4.1.1 Cuestionamiento de modelo

Después que se ha estimado un modelo de regresión se debe, obligatoriamente, estudiar si ese modelo retenido es adecuado. También, se debe probar si supuestos con base en que se sustenta son aceptables a vista de datos. De manera más precisa, ¿será verdad que errores son independientes, normalmente distribuidos con media cero y varianza constante, será lineal el modelo?

No se puede dar una respuesta exacta, pero residuos darán pautas para decidir si no se acepta o no se rechaza estas hipótesis. En primer término, se estudia una prueba de linealidad bajo supuesto que existen observaciones repetidas y, después, se analiza residuos.

4.1.2 Prueba de linealidad

Si para cada upla $(x_{i2}, x_{i3}, x_{i4}, x_{i5}, \dots, x_{ik})$ existen varias observaciones de variable dependiente Y, se puede construir una prueba de linealidad.

4.2 Regresión lineal con observaciones repetidas

Suponga para cada upla $(x_{i2}, x_{i3}, x_{i4}, x_{i5}, \dots, x_{ik})$ se hacen n_i mediciones u observaciones de variable respuesta, así que se tiene el modelo:

$$y_{ir} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_k x_{ik} + u_{ir}\ i = 1,2,3,4,5 \dots, m;\ r = 1,2,3,4,5, \dots, n_i$$

Suponga $\sum_{i=1}^{m} n_i = n, m > k$ y rango de matriz de diseño es k. Las hipótesis respecto a errores son habituales, independientes y normalmente distribuidos con parámetros $(0, \sigma^2)$. Se hallarán estimadores mínimos cuadrados ordinarios (MCO). Sea:

$$f(b_1, b_2, b_3, b_4, \dots, b_k) = \sum_{i=1}^{m}\sum_{r=1}^{n_i}\left(y_{ir} - b_1 - \sum_{j=2}^{k}(b_j x_{ij})\right)^2$$

$$\frac{\partial(f)}{\partial(b_1)}(b_1, b_2, b_3, b_4, \dots, b_k) = -2\sum_{i=1}^{m}\sum_{r=1}^{n_i}\left(y_{ir} - b_1 - \sum_{j=2}^{k}(b_j x_{ij})\right)$$

$$\frac{\partial(f)}{\partial(b_1)}(b_1, b_2, b_3, b_4, \dots, b_k) = -2\sum_{i=1}^{m}\sum_{r=1}^{n_i}\left(y_{ir} - b_1 - \sum_{j=2}^{k}(b_j x_{ij})\right)x_{il}\ \forall l$$

Si se iguala a 0 se tiene el sistema normal de ecuaciones:

$$\begin{cases}\sum_{i=1}^{m}\left[(n_i)\left(\bar{y}_{i.} - \hat{\beta}_1 - \sum_{j=2}^{k}(\hat{\beta}_j x_{ij})\right)\right] = 0 \text{ tal que } \bar{y}_{i.} = \left[\left(\frac{1}{n_i}\right)\left(\sum_{j}(y_{ij})\right)\right] \\ \sum_{i=1}^{m}\left[(n_i)\left(\bar{y}_{i.} - \hat{\beta}_1 - \sum_{j=2}^{k}(\hat{\beta}_j x_{ij})\right)\right]x_{il} = 0\ \forall l\end{cases}$$

Es equivalente a ecuación matricial:

$$(X^t R X)\hat{\beta} = X^t R \bar{Y} \text{ tal que } R = \text{diag}_{(\text{Diagonal})}\{n_1, n_2, n_3, n_4, n_5, \dots, n_k\}$$

$$\bar{Y} = \begin{bmatrix}\bar{y}_{1.}\\ \bar{y}_{2.}\\ \bar{y}_{3.}\\ \bar{y}_{4.}\\ \vdots\\ \bar{y}_{k.}\end{bmatrix}, X = \begin{bmatrix}1 & x_{12} & x_{13} & x_{14} & \cdots & x_{1k}\\ 1 & x_{22} & x_{23} & x_{24} & \cdots & x_{2k}\\ 1 & x_{32} & x_{33} & x_{34} & \cdots & x_{3k}\\ 1 & x_{42} & x_{43} & x_{44} & \cdots & x_{4k}\\ & \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{m2} & x_{m3} & x_{m4} & \cdots & x_{mk}\end{bmatrix}$$

Por hipótesis la matriz X es de rango $k < m$ tal que la matriz $X^t RX$ es de rango completo, tal que:

$$\hat{\beta} = (X^t R X)^{-1}(X^t R \bar{Y})$$

$$\hat{y}_{ir} = \hat{y}_i = (1, x_{i2}, x_{i3}, x_{i4}, x_{i5}, \dots, x_{ik})\hat{\beta} \text{ tal que } r = 1,2,3,4,5, \dots, n_i$$

Prueba. La repetición de mediciones es de suma utilidad para validar el modelo. Para probar el ajuste del modelo $y_{ir} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_k x_{ik} + u_{ir}$ $i = 1,2,3,4,5 \dots, m; r = 1,2,3,4,5, \dots, n_i$ a datos, se compara con modelo:

$$y_{ir} = \mu_i + \varepsilon_{ir} \; i = 1,2,3,4,5 \dots, m; r = 1,2,3,4,5, \dots, n_i$$

Tal que errores $\{\varepsilon_{ir}\}$ son independientes normalmente distribuidos con parámetros $(0, \sigma^2)$. En modelo $y_{ir} = \mu_i + \varepsilon_{ir}$ $i = 1,2,3,4,5 \dots, m; r = 1,2,3,4,5, \dots, n_i$ supone que medias μ_i dependen de valores $(x_{i2}, x_{i3}, x_{i4}, x_{i5}, \dots, x_{ik})$, pero no obedecen a estructura alguna en particular.

Se dice que modelo $y_{ir} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_k x_{ik} + u_{ir}$ $i = 1,2,3,4,5 \dots, m; r = 1,2,3,4,5, \dots, n_i$ es adecuado a nivel de significancia α si hipótesis $H_0: \mu = X\beta$ no es rechazada a nivel α.

Teorema. Bajo hipótesis de normalidad, no se acepta modelo $y_{ir} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_k x_{ik} + u_{ir}$ $i = 1,2,3,4,5 \dots, m; r = 1,2,3,4,5, \dots, n_i$ a favor de modelo $y_{ir} = \mu_i + \varepsilon_{ir}$ $i = 1,2,3,4,5 \dots, m; r = 1,2,3,4,5, \dots, n_i$ a nivel de significancia α si y solo si:

$$\left(\frac{n-m}{m-k}\right) * \left(\frac{\sum_{i=1}^{m}[(n_i)(\bar{y}_{i.} - \hat{y}_{i.})^2]}{\sum_{i=1}^{m}\sum_{r=1}^{n_i}(y_{ir} - \bar{y}_{i.})^2}\right) \geq F_{(m-k;n-m;\alpha)}$$

Demostración. Se debe estimar cociente:

$$F_{(y)} = \left[\left(\frac{n-m}{m-k}\right)\left(\frac{SRC_{(0:\text{ Suma de Residuos al Cuadrado de modelo})} - SRC_{(\text{Suma de Residuos al Cuadrado})}}{SRC_{(\text{Suma de Residuos al Cuadrado})}}\right)\right]$$

Tal que $SRC_{(0:\text{ Suma de Residuos al Cuadrado de modelo})}$ $y_{ir} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_k x_{ik} + u_{ir}$ $i = 1,2,3,4,5 \dots, m; r = 1,2,3,4,5, \dots, n_i$:

$$SRC_{(0:\text{ Suma de Residuos al Cuadrado de modelo})} = \sum_{i=1}^{m}\sum_{r=1}^{n_i}(y_{ir} - \hat{y}_{i.})^2$$

Estimador de mínimos cuadrados de μ_i en modelo $y_{ir} = \mu_i + \varepsilon_{ir}$ $i = 1,2,3,4,5 \dots, m; r = 1,2,3,4,5, \dots, n_i$ es:

$$\tilde{\mu}_i = \bar{y}_{i.}\ i = 1,2,3,4,5 \dots, k;\ SRC_{(\text{Suma de Residuos al Cuadrado})} = \sum_{i=1}^{m}\sum_{r=1}^{n_i}(Y_{ir} - \bar{y}_{i.})^2$$

Vectores $\{Y_{ir} - \bar{y}_{i.}\}$ y $\{\bar{y}_{i.} - \hat{y}_{i.}\}$ son ortogonales:

$$\sum_{i=1}^{m}\sum_{r=1}^{n_i}(y_{ir} - \bar{y}_{i.})(\bar{y}_{i.} - \hat{y}_{i.}) = \left[\sum_{i=1}^{m}(\bar{y}_{i.} - \hat{y}_{i.})\sum_{r=1}^{n_i}(y_{ir} - \bar{y}_{i.})\right] = 0$$

Por teorema de Pitágoras:

$$\|Y_{ir} - \bar{Y}_{i.}\|^2 + \left\|\bar{Y}_{i.} - \hat{Y}_{i.}\right\|^2 = \left\|Y_{ir} - \hat{Y}_{i.}\right\|^2$$

Tal que:

$$SRC_{(0:\ \text{Suma de Residuos al Cuadrado de modelo})} - SRC_{(\text{Suma de Residuos al Cuadrado})}$$
$$= \sum_{i=1}^{m}\sum_{r=1}^{n_i}(y_{ir} - \hat{y}_{i.})^2 - \sum_{i=1}^{m}\sum_{r=1}^{n_i}(y_{ir} - \bar{y}_{i.})^2 = \sum_{i=1}^{m}[(n_i)(\bar{y}_{i.} - \hat{y}_{i.})^2]$$

Definición. Suma $\sum_{i=1}^{m}[(n_i)(\bar{y}_{i.} - \hat{y}_{i.})^2]$ es error por mala adecuación. La suma $\sum_{i=1}^{m}\sum_{r=1}^{n_i}(y_{ir} - \bar{y}_{i.})^2$ es llamado error puro. Resultados se escriben en ANOVA, ADEVA, ANDEVA, ANVA, AVAR o ANOVA de Fisher que puede ser insertada en:

Tabla 24. ANOVA de Fisher

Análisis de Varianza o **ANOVA, ADEVA, ANDEVA, ANVA, AVAR o ANOVA de Fisher**					
F. de V$_{\text{(Factor de Variación)}}$ o **VF**$_{\text{(Variation factor)}}$	**Gl**$_{\text{(Grados de Libertad)}}$ o **Df**$_{\text{(Degrees of freedom)}}$	**SC**$_{\text{(Suma de Cuadrados)}}$ o **Sum Sq**$_{\text{(Sum of square)}}$	**CM**$_{\text{(Cuadrado Medio)}}$ o **Mean Sq**$_{\text{(Mean squares)}}$	**F**$_{\text{(calculada)}}$ o **F value**$_{\text{(Calculated F)}}$	**Pr(> F)** o **ρ − Value**
Error total	$(n-k)$	$\sum_{i=1}^{m}\sum_{r=1}^{n_i}(y_{ir}-\hat{y}_{i.})^2$			
Error por mala adecuación	$(m-k)_{\text{(gl numerador)}}$	$\sum_{i=1}^{m}[(n_i)(\bar{y}_{i.}-\hat{y}_{i.})^2]$	$MMA/gl_{\text{(numerador)}}$	$\frac{MMA_{\text{Adecuación}}}{MEP_{\text{Error}}}$	$\frac{gl_{\text{(numerador)}}}{gl_{\text{(denominador)}}}$
Error puro	$(n-m)_{\text{(gl denominador}}$	$\sum_{i=1}^{m}\sum_{r=1}^{n_i}(y_{ir}-\bar{y}_{i.})^2$	$MEP/gl_{\text{(denominador)}}$		

4.3 Análisis de residuos

Para analizar residuos se recurre a gráficos estudiados, su trazado y uso; aunque, es importante no olvidar graficar errores contra un regresor. Con el objeto de estudiar relación entre regresor X_j y respuesta Y se recurre a gráficos de residuos parciales:

a) Primero se estima Y no explicada por resto de regresores tal que es residuo $\left(\widehat{U}^{(j)}\right)$ de regresión Y sobre $(X_2, X_3, X_4, X_5, \dots, X_k)$.
b) Se grafica puntos $\left(x_{ij}, \hat{u}_i^{(j)}\right)$

La tendencia del gráfico mide efecto marginal de X_j sobre Y. Si regresores $\{X_j\}$ fuesen dos a dos ortogonales estimadores $\hat{\beta}_j$ no cambian si excluye cualesquiera de regresores tal que si no se excluye variables:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + \hat{\beta}_5 x_{i5} + \cdots + \hat{\beta}_k x_{ik} \;\; \hat{u}_i = y_i - \hat{y}_i$$

Si se excluye X_j:

$$\hat{y}_i(j) = \hat{y}_i - \hat{\beta}_j x_{ij}$$

$$\hat{u}_i = y_i - \hat{y}_i(j) = \hat{u}_i - \hat{\beta}_j x_{ij}$$

Tal que si regresores tienen correlación baja se grafica puntos $\left(x_{ij}, \hat{u}_i + \hat{\beta}_j x_{ij}\right)$

4.4 Residuos estandarizados

Es preferible analizar residuos estadarizados, pues ponen en evidencia puntos singulares y normalidad de errores.

Por **Teorema**. Bajo hipótesis H, si matriz X es de rango completo y no aleatoria, 1. $E\left(\widehat{U}\right) = 0$ y 2. $Var\left(\widehat{U}\right) = \sigma^2(I - H)$, $Var\left(\widehat{U}\right) = \sigma^2(I - H)$ tal que $H = XX^t(X^tX)^{-1} \Rightarrow \forall_i, Var(\hat{u}_i) = \sigma^2(1 - h_{ii})$ donde $h_{ii} = x_i x_i^t (X^tX)^{-1}$ es i-ésimo

elemento de diagonal de matriz H tal que se desprende que residuos estandarizados son:

$$r_i = \left[\frac{\hat{u}_i}{\sqrt{s_R^2(1-h_{ii})}}\right]$$

Simplificado sería:

$$\left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} Var(\hat{u}_i)\right) = \left[\left(\frac{1}{n}\right)\sigma^2 tr(I-H)\right] = \left[\frac{(\sigma^2)(n-k)}{n}\right]$$

Esto permite definir residuos normalizados por:

$$r_i^{'} = \left[\sqrt{\left(\frac{n}{n-k}\right)}\left(\frac{\hat{u}_i}{\sqrt{s_R^2}}\right)\right]$$

Si n es grande respecto a k; entonces, la raíz cuadrada es aproximadamente uno tal que algunos textos estandarizan residuos con fórmula:

$$d_i = \left[\frac{\hat{u}_i}{\sqrt{s_R^2}}\right]$$

Si $h_{ii} \approx 0 \Rightarrow r_i \approx d_i$, pero si $h_{ii} \approx 1$ dos residuos serán muy diferentes, r_i es muy grande en términos absolutos respecto a d_i.

4.5 Puntos singulares

La matriz (X^tX) es simétrica definida positiva, h_{ii} se interpreta como norma ecuclideama de vector x_i respecto a $(X^tX)^{-1}$:

$$h_{ii} = \|x_i\|^2 (X^tX)^{-1}$$

Se puede demostrar:

$$h_{ii} = \left(\frac{1}{n}\right)[1 + \tilde{x}_i \tilde{x}_i^t S_{xx}^{-1}]$$

Tal que $\tilde{x}_i = (x_{i2} - \bar{x}_2, x_{i3} - \bar{x}_3, x_{i4} - \bar{x}_4, x_{i5} - \bar{x}_5, \dots, x_{ik} - \bar{x}_k)$, S_{xx} es matriz de varianzas-covarianzas entre columnas de X tal que $(\tilde{x}_i \tilde{x}_i^t S_{xx}^{-1})$ es distancia de Mahalnobis de $(x_{i2}, x_{i3}, x_{i4}, x_{i5}, \dots, x_{ik})$ a centro $(\bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5, \dots, \bar{x}_k)$. Un valor alto de h_{ii} indica que $(x_{i2}, x_{i3}, x_{i4}, x_{i5}, \dots, x_{ik})$ se aleja del centro. Con base en teorema Gauss-Markov se demuestra, entre otras:

$$(\hat{U}) = U(I - H) \Rightarrow u_i - \sum_{j=1}^{n} (h_{ii} u_j)$$

Donde $\hat{u}_i \approx u_i$ ssi $\sum_{j=1}^{n} (h_{ii} u_j) \approx 0$. Si $E(\hat{u}_i) = 0$ y $Var(\sum_{j=1}^{n} (h_{ii} u_j)) = \sigma^2 \sum_{j=1}^{n} (h_{ii}^2) = \sigma^2 h_{ii} \Rightarrow \hat{u}_i \approx u_i$ ssi $h_{ii} \approx 0$, residuo $\hat{u}_i$ es buen indicador de error u_i ssi $h_{ii} \approx 0$.

Se puede demostrar:

$$\left(\frac{1}{n}\right) \leq h_{ii} < 1$$

Tal que, el residuo es informativo sólo cuando $h_{ii} \approx \left(\frac{1}{n}\right)$, si n es grande $h_{ii} \approx 0$. Si $h_{ii} \approx 1$, por $\forall_i, Var(\hat{u}_i) = \sigma^2(1 - h_{ii})$, $Var(\hat{u}_i) \approx 0 \Rightarrow \hat{u}_i \approx 0$ o, equivalente a, $y_i \approx \hat{y}_i$. La superficie de regresión pasará muy cerca del punto (x_i, y_i) sin importar valor de y_i; es decir, es punto palanca.

4.6 Predicción individual

4.6.1 Media

Suponga que se basa en n observaciones tal que se estiman parámetros de una regresión lineal múltiple:

$$Y = X\beta + U$$

Si se ha calculado $\hat{\beta}$ el $EMC_{(\text{Estimador de Mínimos Cuadrados})}$ de β y s_R^2 el estimador insesgado de σ^2. Se está interesado en predecir el valor medio de variable dependiente Y para un vector:

$$x^0 = \left(1, x_2^0, x_3^0, x_4^0, x_5^0, \dots, x_k^0\right)^t$$

Sea Y_0 valor de Y en x^0, $\mu_0 = E(Y_0) = (x^0)^t(\beta)$. Por teorema de Gass-Markov el mejor estimador lineal insesgado de μ_0 es:

$$\widehat{Y}_{(0|n)} = (x^0)^t\left(\hat{\beta}\right)$$

Este valor es predicción de media de μ_0.

Observación. Subíndice n recuerda predicción con base en n observaciones "anteriores a x^0", será válida si la relación funcional entre variables se mantiene más allá del rango observado para valores de predictores $\{X_j\}$.

Teorema. Con hipótesis usuales de mínimos cuadrados ordinarios (MCO) y $rg(X) = k$:

$$\widehat{Y}_{(0|n)} \to N[(x^0)^t\beta, x^0\sigma^2(x^0)^t(X^tX)^t]$$

$$(x^0)^t\left(\hat{\beta}\right) \pm \sqrt{(s^2)(x^0)^t(X^tX)^t(x^0)}t_{\left(n-k;\frac{\alpha}{2}\right)}$$

Este último es intervalo de confianza de nivel de significancia $(1-\alpha)$ para media μ_0.

Teorema. Si vector x_0 es una de las filas $x_{(i)}$ de matriz de datos X tal que:

$$h_{ii} = x_{(i)}x_{(i)}^t(X^tX)^t = \left(\frac{1}{n}\right)\left[1 + (x_i^t - \bar{x})(S_{xx}^{-1})\left(x_{(i)} - \bar{x}\right)\right]$$

Donde S_{xx} es matriz de varianzas-covarianzas, sesgada, de columnas de matriz X, $\bar{x}$ es vector de medias. Producto $(x_i^t - \bar{x})(S_{xx}^{-1})\left(x_{(i)} - \bar{x}\right)$ es distancia Mahalanobis de $x_{(i)}$ a $\bar{x}$ tal que evidencia que varianza $(\sigma^2 h_{ii})$ es mínima cuando $x_{(i)} = \bar{x}$, su valor es $\left(\frac{\sigma^2}{n}\right)$. Para filas muy alejadas de centro de varianza puede aproximarse a $2\left(\frac{\sigma^2}{n}\right)$.

4.6.2 Individual

Valor Y_0 implica error:

$$Y_0 = (x^0)^t\beta + u_0$$

Error u_0 es independiente de errores $(u_1, u_2, u_3, u_4, \dots, u_n)$ corresponden a observaciones, es de ley $N(0, \sigma^2)$. No se puede predecir un valor particular de Y_0, pues error no es predecible, máximo puede predecirse su media y estimar intervalos de confianza para media y predicción. El error de predicción es:

$$e = Y_0 - \hat{Y}_{(0|n)} = (x^0)^t\left(\beta - \hat{\beta}\right) + u_0 \text{ tal que } E(e) = 0, Var(e)$$
$$= \sigma^2(x^0)^t + (x^0)Var\left(\hat{\beta}\right) + \sigma^2$$

Además:

$$e \to N\left[0, \sigma^2\left(x^0(x^0)^tVar\left(\hat{\beta}\right) + 1\right)\right]$$

Tal que, el intervalo de confianza de nivel de significancia $1 - \alpha$ para predicción de un valor:

$$(x^0)^t\hat{\beta} \pm \sqrt{s^2\left[\left(x^0(x^0)^tVar\left(\hat{\beta}\right) + 1\right)\right]}t_{\left(n-k;\frac{\alpha}{2}\right)}$$

4.7 Problemas en Regresión Múltiple

En estimación de parámetros en modelo lineal múltiple se presentan varios problemas a tenerse en cuenta al realizar un ajuste de datos, pues son multicolinealidad, preservación de valores extremos, autocorrelación y no normalidad de errores.

4.7.1 Análisis de errores

Una vez construido el modelo de regresión se comprueba si hipótesis de linealidad, normalidad e independencia se cumplen. La matriz de covarianzas de errores es:

$$\mathrm{Cov}(e) = \sigma^2(I - V) \text{ tal que } V = XX^t(X^tX)^{-1} \Rightarrow \mathrm{Cov}(e_i) = \sigma^2(I - V_{ii})$$

Tal que $V_{(ii)}$ mide distancia entre X_i y $\overline{X}$. Para comparar los residuos suele ser más cómodo cambiarlos de escala, estandarizándolos o estudentizandolos. Los residuos estandarizados se definen:

$$r_i = \left[\frac{e_{(i)}}{s\sqrt{1} - V_{(ii)}}\right]$$

Los residuos estudentizados se estiman mediante:

$$t_i = \left[\frac{e_{(i)}}{s_{(i)}\sqrt{1} - V_{(ii)}}\right]$$

Siguen una ley t con $(n - k - 2)$ grados de libertad tal que $s_{(i)}$ son residuos de regresión cuando se excluye la i -ésima observación.

4.7.2 Error de especificación

Se comete error de especificación cuando se establece una dependencia errónea de respuesta en función de variables explicativas, pues se omiten variables importantes e introduce variables innecesarias o supone una relación lineal cuando la dependencia es no lineal. La especificación incorrecta del modelo conduce a que residuos tengan esperanza no nula y estimadores obtenidos sean sesgados.

4.7.3 No normalidad de residuos

La suposición que residuos ξ están normalmente distribuidos no es necesaria para la estimación de parámetros de regresión ni para participación de variabilidad total. La normalidad es necesaria para la construcción de pruebas de hipótesis e intervalos de confianza sobre los parámetros. El impacto de no normalidad en mínimos cuadrados depende del grado de desviación de no-normalidad y aplicación específica.

Estimadores de parámetros serán insesgados, pero sus intervalos de confianza y pruebas de hipótesis serán incorrectos. Sin embargo, la razón F es razonablemente robusta contra la normalidad. Para detectar la normalidad de errores es conveniente fijarse en coeficientes de asimetría y curtosis. Además, se pueden realizar gráficos $Q - Q$ o $P - P$ de bondad de ajuste a ley normal.

La transformación de variable dependiente a una forma que sea más cercana a normal es recurso muy empleado. Estas transformaciones suelen ser sugeridas por los gráficos de residuos. También, se puede utilizar el método de $Box - Cox$ de transformación potencial. En muchos casos, la desviación de normalidad se debe a presencia de valores atípicos; en otro caso, es conveniente examinar su influencia. Recientemente, se han desarrollado modelos que consideran errores están distribuidos separados por una ley t, de un numero de grados de libertad desconocido, como una generalización de hipótesis de normalidad.

4.7.4 Puntos atípicos o inusuales

Se espera que datos correspondientes a observaciones se encuentren distribuidos en una región más o menos cercana, pero puede suceder que una o varias observaciones estén alejadas del resto de los datos. Estas observaciones pueden fluir mucho en modelo final. Se puede disponer de 100 observaciones y construir con ellos un modelo con propiedades debidas únicamente a dos puntos. Conocer si este tipo de puntos influye perjudicialmente en modelo permite mejorarlo.

La primera forma para determinar si un valor es atípico es mediante los residuos estudentizados. Se comparan valores de t_i con valores críticos de una ley t con $(n - k - 2)$ grados de libertad. Otra forma de conocer cuáles son puntos distantes es mediante distancia:

$$D_i^2 = \sum_{j=1}^{k} \left(\frac{b_j \, (x_{ij} \bar{x}_j}{\sqrt{MCE}} \right)^2 \; i = 1,2,3,4,\ldots,n$$

El procedimiento consiste en ajustar el modelo, estimar $D_i^2, i = 1,2,3,4,\ldots,n$, y ordenarlas en forma ascendente de acuerdo con el D_i^2. Los puntos con alto D_i^2 son inusuales. Existen otros tipos de distancia que permiten la detección de valores atípicos y puntos influyentes a cada uno con distintas propiedades, pero todas siguen igual principio para identificar tales observaciones.

Identificación. Para la detección de la autocorrelación se emplea el estadístico de Durbin-Watson:

$$r_h = \frac{\sum e_i e_{i-}h}{\sum e_i^2}$$

Para la realización de pruebas estadísticas sobre la autocorrelación existen tablas, pero dar la siguiente regla empírica para detectar la autocorrelación de orden 1. Se construye:

$$d = 2(1 - r_1)$$

Si su valor es próximo a 2 no existirá autocorrelación, pero si d tiene un valor entre 0 y 2 será positiva y si el valor d esta entre 2 y 4 indica que correlación será negativa. En los programas informáticos se suele calcular su valor y el nivel de simplificación de la prueba. Alternativamente, se aplica el estadístico Box – Ljung para detectar la autocorrelación.

Tratamiento. Para explicar la evolución de variables que tiene un comportamiento en que aparece autocorrelación es conveniente usar métodos de análisis de series de tiempo, que permiten abordar de mantenimiento global el problema de construcción de modelos para estas variables. También, se pueden utilizar métodos especiales de regresión, como los mínimos cuadrados generalizados o modelos generalizados.

Por la dificultad que entraña la realización manual de cálculos; especialmente en determinación de potenciales problemas que modelo pudiera presentar. Se hace mediante el empleo de programas estadísticos especializados, que facilitan su cálculo, correspondiendo al usuario la interpretación correcta de los resultados. Deberá notar que temas tratados sólo cubren la parte central de análisis de regresión.

4.7.5 Multicolinealidad

El término multicolinealidad se atribuye a Ragnar Frisch, que asigna una relación lineal "perfecta" o exacta entre algunas o todas las variables explicativas de un modelo de regresión. En regresión con k variables que incluye las variables explicativas $(X_1, X_2, X_3, X_4, \dots, X_k)$ tal que existe una relación lineal exacta si se satisface la siguiente condición:

$$\lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \lambda_4 X_4 + \cdots + \lambda_k X_k$$

Multicolinealidad incluye el caso de multicolinealidad perfecta, como ecuación anterior y, también, caso en que hay X variables intercorrelacionadas, pero no en forma perfecta:

$$\lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \lambda_4 X_4 + \cdots + \lambda_k X_k + \nu_{i\ (\text{Término de error estocástico})} = 0$$

Para apreciar la diferencia entre multicolinealidades perfecta y menos que perfecta suponga; por ejemplo: $\lambda_2 \neq 0$:

$$X_{2i} = -\left[X_{1i}\left(\frac{\lambda_1}{\lambda_2}\right)\right] - \left[X_{3i}\left(\frac{\lambda_3}{\lambda_2}\right)\right] - \left[X_{4i}\left(\frac{\lambda_4}{\lambda_2}\right)\right] - \left[X_{5i}\left(\frac{\lambda_{15}}{\lambda_2}\right)\right] - \cdots - \left[X_{ki}\left(\frac{\lambda_k}{\lambda_2}\right)\right]$$

Muestra la forma como X_2 está exactamente relacionada de manera lineal con otras variables, o cómo se deriva de una combinación lineal de otras variables X.

El coeficiente de correlación entre variable X_2 y combinación lineal del lado derecho de ecuación pasada está obligado a ser igual a uno. Paralelamente, si

$\lambda_2 \neq 0$, ecuación $\lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \lambda_4 X_4 + \cdots + \lambda_k X_k +$ $\nu_{i\ (\text{Término de error estocástico})} = 0$ se escribe:

$$X_{2i} = -\left[X_{1i}\left(\frac{\lambda_1}{\lambda_2}\right)\right] - \left[X_{3i}\left(\frac{\lambda_3}{\lambda_2}\right)\right] - \left[X_{4i}\left(\frac{\lambda_4}{\lambda_2}\right)\right] - \left[X_{5i}\left(\frac{\lambda_{15}}{\lambda_2}\right)\right] - \cdots$$
$$- \left[X_{ki}\left(\frac{\lambda_k}{\lambda_2}\right) - \nu_{i\ (\text{Término de error estocástico})}\left(\frac{1}{\lambda_2}\right)\right]$$

Esto demuestra que X_2 no es combinación lineal exacta de otras X, pues está determinada por $\nu_{i\ (\text{Término de error estocástico})}$. El método algebraico para problema de multicolinealidad se expresa concisamente mediante un diagrama de Ballentine:

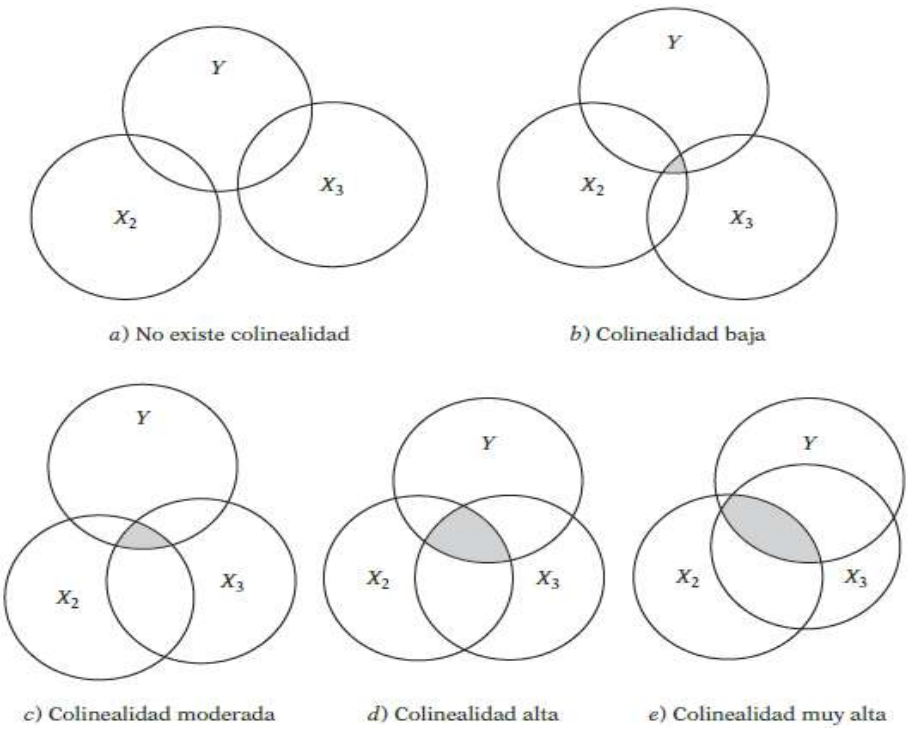

Figura 6. *Medición del grado de colinealidad mediante diagrama de Ballentine*

En otras figuras:

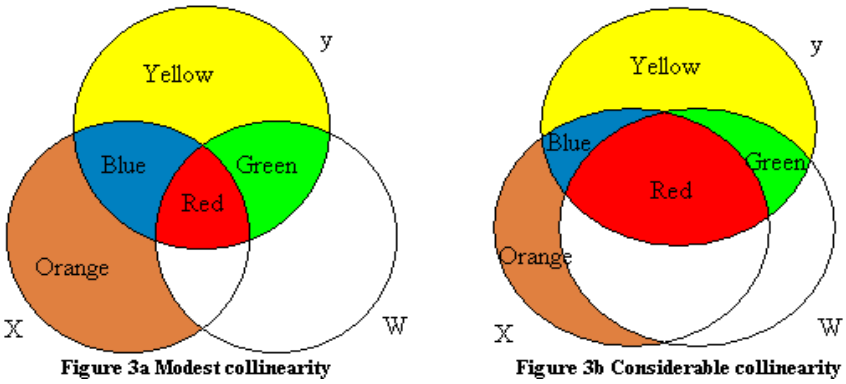

Figura 7. *Medición del grado de colinealidad mediante diagrama de Ballentine*

Los círculos Y, X_2 y X_3 indican variaciones en $Y_{(\text{variable dependiente})}$ en X_2 y X_3, variables explicativas.

El grado de colinealidad se mide por magnitud de intersección, área sombreada, de círculos X_2 y X_3. En figuras 10.1a), no hay intersección entre X_2 y X_3 tal que no hay colinealidad; de 3.3.b) a 3.3.e), el grado de colinealidad va de "bajo" a "alto", pues entre mayor sea la intersección entre X_2 y X_3 -entre mayor sea el área sombreada- mayor será el grado de colinealidad. Si X_2 y X_3 estuvieran superpuestos completamente o si X_2 estuviera por completo dentro de X_3, o viceversa, la colinealidad sería perfecta.

Multicolinealidad, como se define, refiere sólo a relaciones lineales entre las variables X tal que no aplica a las relaciones no lineales entre ellas:

$$Y_{i\,(\text{Costo de producción total})} = \beta_1 + \beta_2 X_{i\,(\text{Producción})} + \beta_3 X^2_{i\,(\text{Producción al cuadrado})} + \beta_4 X^3_{i\,(\text{Producción al cubo})} + u_{i\,(\text{Término de error estocástico})}$$

Modelos como estos, no violan el supuesto de no multicolinealidad. Sin embargo, en aplicaciones concretas, el coeficiente de correlación medido de forma convencional demostrará que $X_{i\,(\text{Producción})}, X^2_{i\,(\text{Producción al cuadrado})}$ y $X^3_{i\,(\text{Producción al cubo})}$ están altamente correlacionadas, que dificultará estimar

parámetros de ecuación anterior con mayor precisión o errores estándar pequeños.

El modelo clásico de regresión lineal supone no hay multicolinealidad entre X_i, pues si la multicolinealidad es perfecta en sentido de $\lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \lambda_4 X_4 + \cdots + \lambda_k X_k$ tal que coeficientes de regresión de variables X son indeterminados y errores estándar infinitos.

Si multicolinealidad es menos que perfecta, como $\lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \lambda_4 X_4 + \cdots + \lambda_k X_k + \nu_{i\ (\text{Término de error estocástico})} = 0$, coeficientes de regresión; aunque sean determinados, poseen grandes errores estándar en relación con coeficientes mismos, que significa que coeficientes no pueden ser estimados con gran precisión o exactitud.

La multicolinealidad puede ser por factores:

a) **Método de recolección de información**. Por ejemplo, la obtención de muestras en un intervalo limitado de valores tomados por regresoras en población.
b) **Restricciones en modelo o población objeto de muestreo**. Por ejemplo, en regresión del consumo de electricidad sobre ingreso (X_2) y tamaño de viviendas (X_3) hay una restricción física en población, pues familias con ingresos más altos suelen habitar viviendas más grandes que familias con ingresos más bajos.
c) **Especificación del modelo**. Por ejemplo, adición de términos polinomiales a un modelo de regresión, específicamente cuando rango de la variable X es pequeño.
d) **Modelo sobredeterminado**. Esto sucede cuando el modelo tiene más variables explicativas que número de observaciones. Esto puede suceder en investigación médica, en ocasiones hay un número reducido de pacientes sobre quienes se reúne información respecto a gran número de variables.

e) **Regresoras de modelo compartan una tendencia común**. Datos de series de tiempo aumenten o disminuyan en el tiempo tal que en regresión del gasto de consumo sobre el ingreso, riqueza y población regresoras, ingreso, riqueza y población quizás todas crezcan con el tiempo a una tasa aproximadamente igual; por lo que, presentaría colinealidad entre dichas variables.

Si se satisfacen supuestos del modelo clásico, estimadores de MCO de coeficientes de regresión son MELI o MEI si se añade el supuesto de normalidad. Puede demostrarse que, aunque la multicolinealidad sea muy alta, como en el caso de casi multicolinealidad, los estimadores de MCO conservarán la propiedad MELI.

En los casos de casi o alta multicolinealidad es probable que se presenten consecuencias:

a) Estimadores de MCO son MELI, presentan varianzas y covarianzas grandes que dificultan estimación precisa.

Para ver varianzas y covarianzas grandes, para modelo $y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i$, varianzas, covarianzas de $\hat{\beta}_2$ y $\hat{\beta}_3$ están dadas por:

$$\mathrm{Var}(\hat{\beta}_2) = \left[\frac{\sigma^2}{\sum\left[(x_{2i}^2)\left(1 - r^2_{23(\text{Coeficiente de correlación entre } x_2 - x_3)}\right)\right]}\right]$$

$$\mathrm{Var}(\hat{\beta}_3) = \left[\frac{\sigma^2}{\sum\left[(x_{3i}^2)\left(1 - r^2_{23(\text{Coeficiente de correlación entre } x_2 - x_3)}\right)\right]}\right]$$

$$\mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_3) = \left[\frac{\left(-r^2_{23(\text{Coeficiente de correlación entre } x_2 - x_3)}\right)(\sigma^2)}{\left(\sqrt{\sum[(x_{2i}^2)(x_{3i}^2)]}\right)\left(1 - r^2_{23(\text{Coeficiente de correlación entre } x_2 - x_3)}\right)}\right]$$

Con base en estas ecuaciones, se desprende que, a medida que r_{23} tiende a 1 tal que a medida que aumenta la colinealidad lo hacen varianzas de estimadores

y, en el límite, cuando $r_{23} = 1$, son infinitas. En caso de última ecuación, a medida que r_{23} aumenta a 1, la covarianza de dos estimadores aumenta en valor absoluto tal que $\text{Cov}(\hat{\beta}_2, \hat{\beta}_3) = \text{Cov}(\hat{\beta}_3, \hat{\beta}_2)$.

La velocidad con que se incrementan varianzas y covarianzas se mide mediante (FIV):

$$\text{FIV}_{(\text{Factor inflacionario})} = \left[\frac{1}{\left(1 - r^2_{23(\text{Coeficiente de correlación entre } x_2 - x_3)}\right)}\right]$$

$\text{FIV}_{(\text{Factor inflacionario})}$ muestra la forma como la varianza de un estimador se infla por presencia de la multicolinealidad, pues a medida que $r^2_{23(\text{Coeficiente de correlación entre } x_2 - x_3)}$ se acerca a 2, $\text{FIV}_{(\text{Factor inflacionario})}$ se acerca a infinito tal que medida que grado de colinealidad aumenta varianza de un estimador también y, en límite, se vuelve infinita.

Si no existe colinealidad entre X_2 y X_3, $\text{FIV}_{(\text{Factor inflacionario})}$ será 1 tal que con esta definición, $\text{Var}(\hat{\beta}_2) = \left[\frac{\sigma^2}{\sum\left[(x^2_{2i})\left(1 - r^2_{23(\text{Coeficiente de correlación entre } x_2 - x_3)}\right)\right]}\right]$ y $\text{Var}(\hat{\beta}_3) = \left[\frac{\sigma^2}{\sum\left[(x^2_{3i})\left(1 - r^2_{23(\text{Coeficiente de correlación entre } x_2 - x_3)}\right)\right]}\right]$, será $\text{FIV}_{(\text{Factor inflacionario})} = 1$:

$$\text{Var}(\hat{\beta}_2) = \left[\left(\frac{\sigma^2}{\sum[(x^2_{2i})]}\right)\left(\text{FIV}_{(\text{Factor inflacionario})}\right)\right]$$

$$\text{Var}(\hat{\beta}_3) = \left[\left(\frac{\sigma^2}{\sum[(x^2_{3i})]}\right)\left(\text{FIV}_{(\text{Factor inflacionario})}\right)\right]$$

Esto demuestra que varianzas $\hat{\beta}_2$ y $\hat{\beta}_3$ son directamente proporcionales a $\text{FIV}_{(\text{Factor inflacionario})}$. Si se extiende a modelo con k variables:

$$\text{Var}(\hat{\beta}_j) = \left[\left(\frac{\sigma^2}{\sum[(x^2_j)]}\right)\left(\frac{1}{1 - R^2_j}\right)\right]$$

b) Con base en consecuencia anterior, intervalos de confianza tienden a ser mucho más amplios, propicia una aceptación más fácil de "hipótesis nula cero"; es decir, el verdadero coeficiente poblacional es cero.

Con base en errores estándar grandes, los intervalos de confianza para los parámetros poblacionales relevantes tienden a ser mayores:

Tabla 25. *Intervalos de confianza para los parámetros poblacionales relevantes*

Efecto de incrementar la colinealidad sobre intervalo de confianza a 95% para $\boldsymbol{\beta_2}$: $\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}} \pm \mathbf{1.96\ ee}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}})$	
Valor $r^2_{23(\text{Coeficiente de correlación entre } x_2 - x_3)}$	Intervalos de confianza a 95% para $\boldsymbol{\beta_2}$
0.00	$\hat{\beta}_2 \pm 1.96 \sqrt{\left(\frac{\sigma^2}{\sum[(x_j^2)]}\right)}$
0.50	$\hat{\beta}_2 \pm 1.96\sqrt{(1.33)} \sqrt{\left(\frac{\sigma^2}{\sum[(x_j^2)]}\right)}$
0.95	$\hat{\beta}_2 \pm 1.96\sqrt{(10.26)} \sqrt{\left(\frac{\sigma^2}{\sum[(x_j^2)]}\right)}$
0.995	$\hat{\beta}_2 \pm 1.96\sqrt{(100.00)} \sqrt{\left(\frac{\sigma^2}{\sum[(x_j^2)]}\right)}$
0.999	$\hat{\beta}_2 \pm 1.96\sqrt{(500.00)} \sqrt{\left(\frac{\sigma^2}{\sum[(x_j^2)]}\right)}$

En casos de alta multicolinealidad, datos muestrales pueden ser compatibles con un diverso conjunto de hipótesis. De ahí que aumente la probabilidad de aceptar hipótesis falsa; es decir, error tipo II.

c) Con base en consecuencia primera, razón t de uno o más coeficientes tiende a ser estadísticamente no significativa.

Para probar hipótesis nula que, por ejemplo $H_0: \hat{\beta}_2 = 0$ se usa razón t tal que $t_{(Calculado)} = \left[\frac{\hat{\beta}_2}{ee(\hat{\beta}_2)}\right]$ vs $t_{(Tablas)}$; aunque, en casos de alta colinealidad, errores estándar estimados aumentan drásticamente, que disminuye los valores t.

Por consiguiente, no se rechaza cada vez con mayor facilidad H_0, que el verdadero valor poblacional relevante es cero .

d) Aún con razón t de uno o más coeficientes sea estadísticamente no significativa, R^2, la medida global de bondad de ajuste, puede ser muy alta.

Si modelo de regresión lineal con k variables es:

$$Y_{i\ (\text{Costo de producción total})} = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \cdots + \beta_k X_{ki} + u_{i\ (\text{Término de error estocástico})}$$

En casos de alta colinealidad es posible encontrar que uno o más coeficientes parciales de pendiente son, de manera individual, no significativos estadísticamente con base en la prueba t. Aun así, R^2 puede ser tan alto, quizás superior a 0.9, que, con base en razón F, es posible rechazar convincentemente hipótesis $H_0: \beta_2 = \beta_3 = \beta_4 = \cdots = \beta_k = 0$. Esta es una de las señales de multicolinealidad, pues valores t son no significativos, pero tiene valor R^2 global alto y razón F significativo.

e) Estimadores de MCO y sus errores estándar son sensibles a pequeños cambios en datos.

Si multicolinealidad no es perfecta, es posible estimación de coeficientes de regresión; sin embargo, sus estimaciones y errores estándar se tornan muy sensibles al más ligero cambio de datos. En presencia de una alta colinealidad, no se pueden estimar los coeficientes de regresión individuales en forma precisa, pero sus combinaciones lineales se estiman con mayor exactitud.

4.7.6 Autocorrelación

Es "correlación entre miembros de series de observaciones ordenadas en tiempo, datos de series de tiempo, o en espacio, datos de corte transversal". Autocorrelación o dependencia secuencial es una herramienta estadística usada frecuentemente en procesado de señales. La función de autocorrelación se define como la correlación cruzada de señal consigo misma. La función de autocorrelación resulta de gran utilidad para encontrar patrones repetitivos dentro de una señal, como periodicidad de señal enmascarada bajo el ruido o para identificar la frecuencia fundamental de una señal que no contiene dicha componente, pero aparecen numerosas frecuencias armónicas de esta. Los procesos de raíz unitaria, autorregresivos, tendencia estacionaria y Modelos de Medias Móviles son ejemplos de procesos con autocorrelación. En contexto de regresión, el modelo clásico de regresión lineal supone que no existe tal autocorrelación en perturbaciones $u_{i\ (\text{Término de error estocástico})}$:

$$\text{Cov}\left(u_{i\ (\text{Término de error estocástico})}, u_{j\ (\text{Término de error estocástico})} \middle| x_i, x_j\right) = \text{E}\left(u_{i\ (\text{Término de error estocástico})}, u_{i\ (\text{Término de error estocástico})}\right) = 0 \text{ i} \neq \text{j}$$

El modelo clásico supone que el término de perturbación relacionado con una observación cualquiera no recibe influencia del término de perturbación relacionado con cualquier otra observación. Por ejemplo: si existe información productiva trimestral de series de tiempo que es inferior en este ciclo, no hay razón para esperar que sea baja en siguiente periodo. En forma similar, si se trata con información de corte transversal que implica la regresión del gasto de consumo familiar sobre el ingreso familiar, no se esperaría que efecto de incremento en ingreso de una familia sobre su gasto de consumo incida en gasto de consumo de otra.

Además, si se trata con información de corte transversal que implica la regresión del gasto de consumo familiar sobre ingreso familiar, no esperaría

que efecto de un incremento en ingreso de una familia sobre su gasto de consumo incida en gasto de consumo de otra. No obstante, si existe tal dependencia, hay autocorrelación:

$$E(u_{i\ (\text{Término de error estocástico})}, u_{j\ (\text{Término de error estocástico})}) \neq 0\ i \neq j$$

Tal que la interrupción ocasionada por una huelga este trimestre puede afectar muy fácilmente la producción del siguiente trimestre o incrementos del gasto de consumo de una familia pueden muy bien inducir a otra familia a aumentar su gasto de consumo para no quedar rezagada. La correlación serial sucede por varias razones:

a) **Incercia o pasividad.**

Las series de tiempo como PNB, índices de precios, producción, empleo y desempleo presentan ciclos económicos. A partir del fondo de la recesión, cuando se inicia la recuperación económica, la mayoría de estas series empieza a moverse hacia arriba. En este movimiento ascendente, el valor de una serie en un punto del tiempo es mayor que su valor anterior. Se genera un "impulso" en ellas y continuará hasta que suceda otra cosa. Por ejemplo: un aumento en tasa de interés, impuestos o ambos para reducirlo. Por consiguiente, es probable que en regresiones que consideran datos de series de tiempo, las observaciones sucesivas sean interdependientes.

b) **Sesgo de especificación: caso de variables excluidas.**

Con frecuencia el investigador empieza con un modelo de regresión razonable que puede no ser "perfecto". Después del análisis de regresión, el investigador haría el examen post mortem para ver si los resultados coinciden con las expectativas a priori. Por ejemplo, el investigador graficaría residuos $\hat{u}_{i\ (\text{Término de error estocástico})}$ obtenidos de regresión ajustada y observaría patrones:

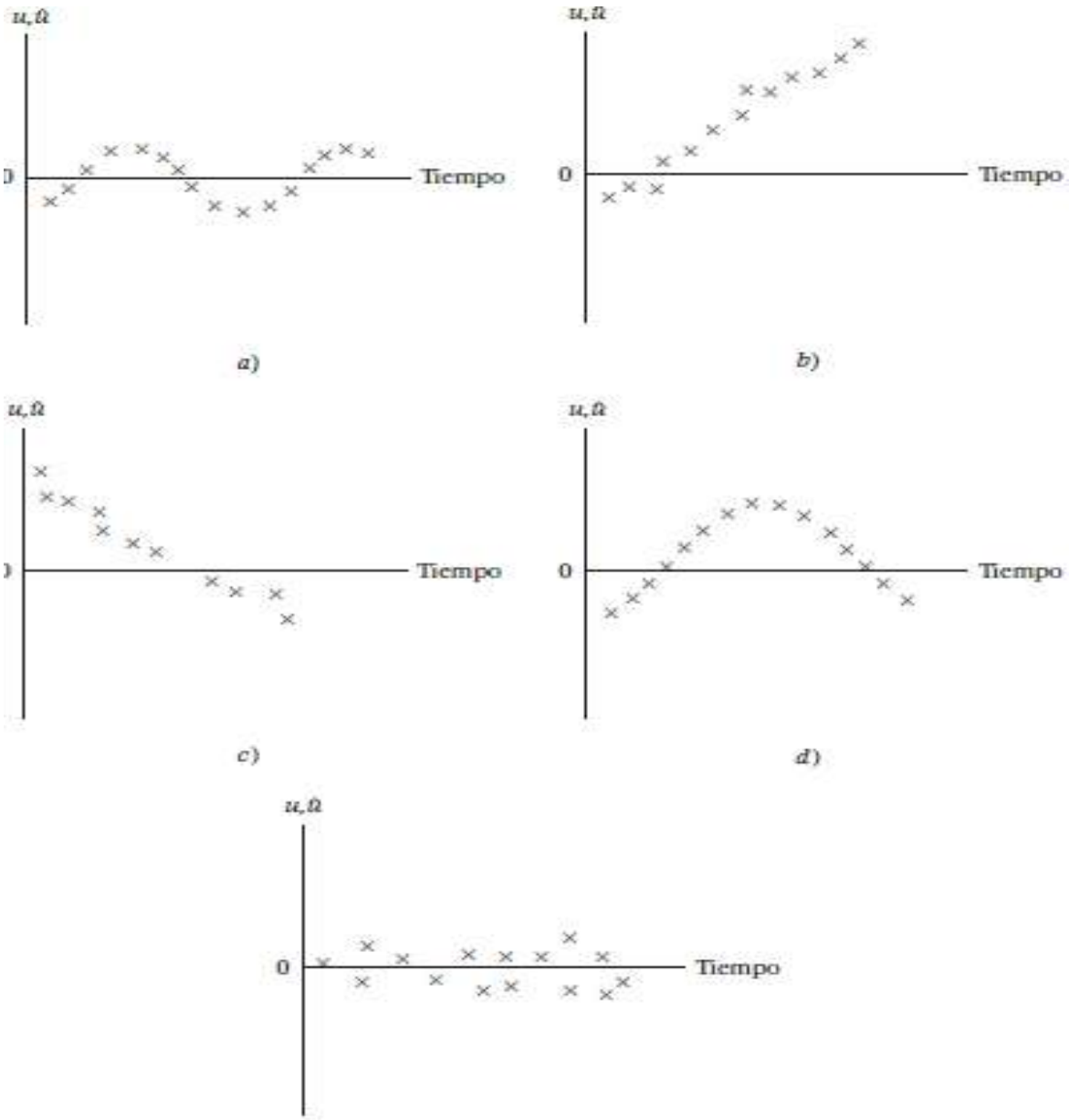

Figura 8. *Gráfica de residuos* $\hat{u}_{i\ (Término\ de\ error\ estocástico)}$ *obtenidos de la regresión ajustada y la observación de patrones*

Estos residuos, representaciones de $u_{i\ (Término\ de\ error\ estocástico)}$, pueden sugerir la inclusión de algunas variables originalmente candidatas pero que no se incluyeron en el modelo por diversas razones. Es el caso del sesgo de especificación ocasionado por variables excluidas. Con frecuencia, la inclusión de tales variables elimina patrón de correlación observado entre los residuales.

Por ejemplo:

$$Y_{i\ (\text{Cantidad de res demandada})} = \beta_1 + \beta_2 X_{2t\ (\text{Precio de carne de res y t tiempo})} + \beta_3 X_{3t\ (\text{Ingreso de consumidor y t tiempo})} + \beta_4 X_{4t\ (\text{Precio del cerdo y t tiempo})} + u_{i\ (\text{Término de error estocástico})}$$

No obstante, se efectúa por alguna razón:

$$Y_{i\ (\text{Cantidad de res demandada})} = \beta_1 + \beta_2 X_{2t\ (\text{Precio de carne de res y t tiempo})} + \beta_3 X_{3t\ (\text{Ingreso de consumidor y t tiempo})} + \nu_{t\ (\text{Término de error estocástico})}$$

Si el primer modelo es "correcto", el "verdadero" o relación auténtica es el segundo modelo tal que equivale a permitir $\nu_{t\ (\text{Término de error estocástico})} = \beta_4 X_{4t\ (\text{Precio del cerdo y t tiempo})}$; por lo tanto, en la medida en que el precio del cerdo afecte el consumo de carne de res, el término de error o de perturbación $\nu_{t\ (\text{Término de error estocástico})}$ reflejará un patrón sistemático que crea una falsa autocorrelación.

Una prueba sencilla de esto sería llevar a cabo primer modelo y, también, segundo modelo tal que se vea si la autocorrelación observada en segundo modelo, de existir, desaparece cuando se efectúa el primero.

c) **Sesgo de especificación: forma funcional incorrecta**.

Suponga que modelo "verdadero" o correcto en un estudio de costo-producción es:

$$Y_{i\ (\text{Costo marginal})} = \beta_1 + \beta_2 X_{i\ (\text{Producción})} + \beta_3 X^2_{i\ (\text{Producción al cuadrado})} + u_{i\ (\text{Término de error estocástico})}$$

Este modelo ajustado es:

$$Y_{i\ (\text{Costo marginal})} = \alpha_1 + \alpha_2 X_{i\ (\text{Producción})} + u_{i\ (\text{Término de error estocástico})}$$

Tal que la curva de costo marginal que corresponde al "verdadero" modelo con curva de costo lineal "incorrecta" es:

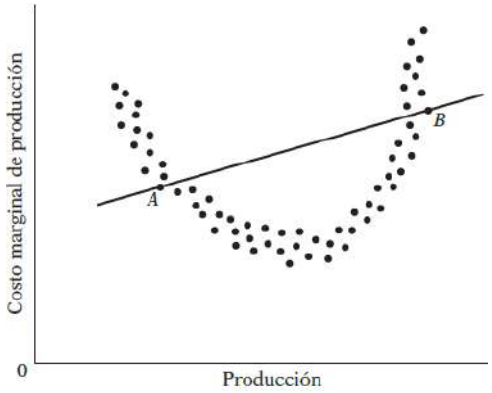

Figura 9. *Curva de costo marginal que corresponde al "verdadero" modelo con curva de costo lineal "incorrecta"*

Entre puntos A y B de curva de costo marginal lineal sobreestimará consistentemente el costo marginal verdadero, mientras que más allá de estos puntos, lo subestimará consistentemente. Este resultado es de esperarse porque el término de perturbación $\nu_{i\ (\text{Término de error estocástico})} = u_{i\ (\text{Término de error estocástico})}$ tal que capta el efecto sistemático del término producción sobre el costo marginal. $\nu_{i\ (\text{Término de error estocástico})}$ reflejará autocorrelación por el uso de una forma funcional incorrecta.

d) **Fenómenos de telaraña.**

La oferta de muchos productos agrícolas refleja el llamado fenómeno de telaraña, en donde la oferta reacciona al precio con un rezago de un periodo, pues debido a instrumentación de decisiones de oferta tarda algún tiempo, llamado periodo de gestación. Por tanto, en siembra de cultivos al principio de año, agricultores reciben influencia del precio prevaleciente el año pasado tal que su función de oferta es:

$$Y_{t\ (\text{Oferta t})} = \beta_1 + \beta_2 P_{(t-1)\ (\text{Producción rezagada})} + u_{t\ (\text{Término de error estocástico})}$$

Suponga que a final de periodo t, el precio $P_{(t)\ (\text{Producción actual})}$ es inferior a $P_{(t-1)\ (\text{Producción rezagada})}$. En consecuencia, es muy probable que en periodo $(t+1)$ agricultores decidan producir menos de lo producido en periodo t. En esta situación no esperaría que perturbaciones $u_{i\ (\text{Término de error estocástico})}$ sean aleatorias, pues si agricultores producen excedentes en el año t, es probable que reduzcan su producción en $(t+1)$ y, sucesivamente, para generar un patrón de telaraña.

e) **Rezagos.**

En una regresión de series de tiempo del gasto de consumo sobre ingreso no es raro hallar que gasto de consumo en periodo actual dependa, entre otras cosas, del gasto de consumo del periodo anterior:

$$Y_{t\ (\text{Consumo t})} = \beta_1 + \beta_2 P_{(t)\ (\text{Ingreso t})} + \beta_3 C_{(t-1)\ (\text{Consumo rezagado})} + u_{t\ (\text{Término de error estocástico})}$$

Esta función es autorregresiva, pues es una variable explicativa es el valor rezagado de la variable dependiente. Los consumidores no cambian sus hábitos de consumo fácilmente por razones psicológicas, tecnológicas o institucionales. Si se ignora el término $C_{(t-1)\ (\text{Consumo rezagado})}$, $u_{t\ (\text{Término de error estocástico})}$ resultante reflejará un patrón sistemático debido a influencia del consumo rezagado en consumo actual.

f) **"Manipulación" de datos.**

En análisis empírico con frecuencia se "manipulan" datos simples. Por ejemplo: en regresiones de series de tiempo con datos trimestrales, por lo general provienen de datos mensuales a los que se agregan simplemente observaciones de tres meses y se divide entre 3. Este procedimiento de promediar cifras suaviza en cierto grado datos al eliminar fluctuaciones en datos mensuales.

Por consiguiente, la gráfica referente a datos trimestrales aparece mucho más suave que la que contiene datos mensuales tal que este suavizamiento puede, por sí mismo, inducir un patrón sistemático en perturbaciones, lo que agrega autocorrelación. Otra fuente de manipulación es interpolación o extrapolación de datos. Estas técnicas de "manejo" podrían imponer sobre datos un patrón sistemático que quizá no estaría presente en datos originales

g) **Transformación de datos.**

Considere:

$$Y_{t\ (\text{Gasto de onsumo t})} = \beta_1 + \beta_2 X_{(t)\ (\text{Ingreso t})} + u_{t\ (\text{Término de error estocástico})}$$

Esta ecuación es válida para cada periodo y, también, es para el periodo anterior $(t - 1)$. Así, ecuación forma de nivel es:

$$Y_{t-1\ (\text{Consumo rezagado t}-1)} = \beta_1 + \beta_2 X_{(t-1)\ (\text{Ingreso rezagado t}-1)} + u_{t-1\ (\text{Término de error estocástico rezagado t}-1)}$$

En este caso están rezagados un periodo. Si restan las ecuaciones anteriores se obtiene ecuación forma en primeras diferencias:

$$\Delta_{(\text{Operador de primeras diferencias})} Y_{t\ (\text{Incremento de consumo t})} = \beta_2 \Delta_{(\text{Operador de primeras diferencias})} X_{(t)\ (\text{Incremento de ingreso t})} + \Delta_{(\text{Operador de primeras diferencias})} u_{t\ (\text{Incremento de término de error estocástico})}$$

Tal que, el modelo dinámico de regresión, modelo con regresadas rezagadas, es:

$$\Delta_{(\text{Operador de primeras diferencias})} Y_{t\ (\text{Incremento de consumo t})} = \beta_2 \Delta_{(\text{Operador de primeras diferencias})} X_{(t)\ (\text{Incremento de ingreso t})} + \Delta_{(\text{Operador de primeras diferencias})} \nu_{t\ (\text{Incremento de término de error estocástico})}$$

Si $Y_{t\ (\text{Gasto de onsumo t})} = \beta_1 + \beta_2 X_{(t)\ (\text{Ingreso t})} + u_{t\ (\text{Término de error estocástico})}$ satisface supuestos usuales de MCO, específicamente inexistencia de autocorrelación, se prueba probar que término de error $v_{t\ (\text{Incremento de término de error estocástico})}$ en última ecuación está autocorrelacionado.

h) **No estacionalidad.**

Al trabajar con datos de series de tiempo, quizá habría que averiguar si una determinada serie de tiempo es estacionaria. De manera informal, una serie de tiempo es estacionaria si sus características, como media, varianza y covarianza, son invariantes respecto del tiempo; es decir, no cambian en relación con el tiempo. Si no es así, tenemos una serie de tiempo no estacionaria.

En un modelo de regresión como $Y_{i\ (\text{Costo marginal})} = \alpha_1 + \alpha_2 X_{i\ (\text{Producción})} + u_{i\ (\text{Término de error estocástico})}$ es muy probable que Y, X sean no estacionarias y, por consiguiente, que el error u también sea no estacionario. En ese caso, el término de error mostrará autocorrelación. En conclusión, existen varias razones por las que término de error en un modelo de regresión pueda estar autocorrelacionado. Autocorrelación puede ser positiva, o negativa; aunque, por lo general, la mayoría de las series de tiempo económicas muestra autocorrelación positiva, pues casi todas se desplazan hacia arriba o hacia abajo en extensos periodos y no exhiben un movimiento ascendente y descendente constante:

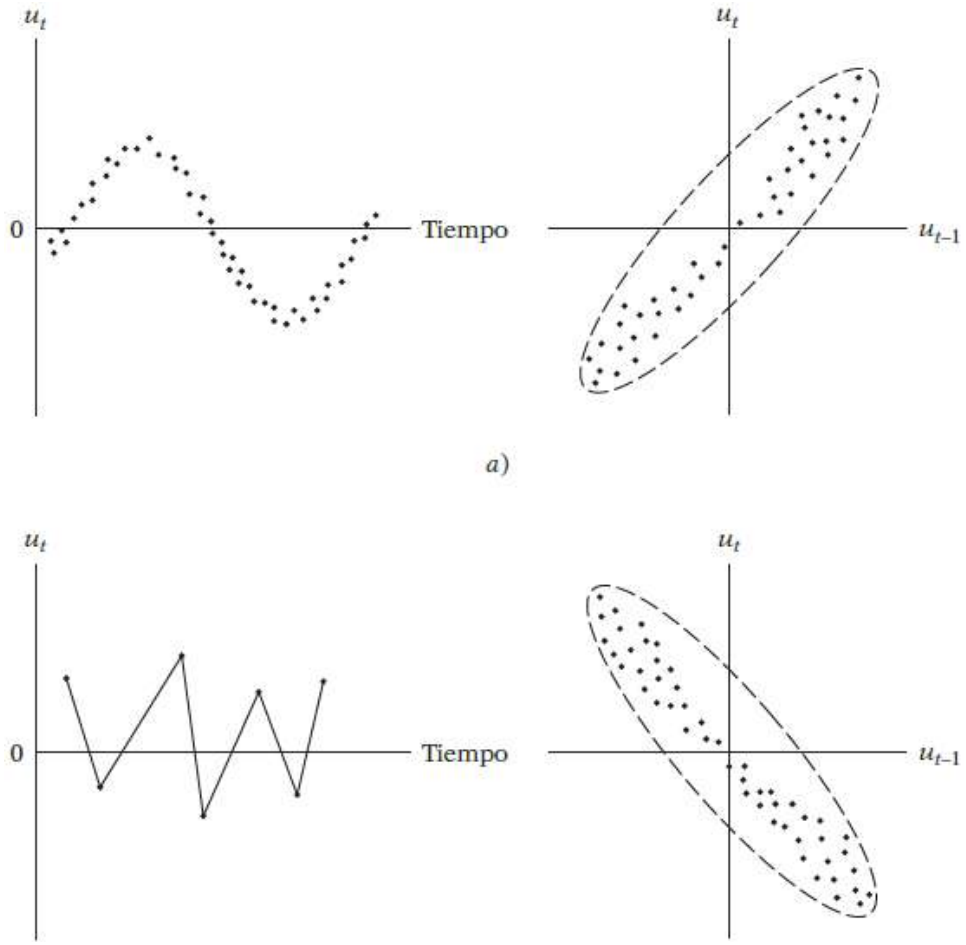

Figura 10. *Movimiento ascendente y descendente constante de la Autocorrelación*

BIBLIOGRAFÍA

Águila, L. M. (2012). *Caracterización morfológica y sensorial del cacao nacional (Theobroma cacao L) a nivel de fincas en Cantón Las Naves, Provincia Bolívar.* Guaranda, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Aguirre, Q. V. (2015). *Identificación de cadenas de comercialización de arroz (Oriza sativa L) en pequeños productores de "Cooperativa Alianza Definitiva" de Cuenca Central de Río Daule, Provincia del Guayas.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecurias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Álava, V. L. (2013). *Respuesta del cultivo de banano (Musa sapientum) a aplicación de dos tipos de fertilizantes foliares orgánicos en diferentes fases lunares en Cantón Pueblo viejo, Provincia Los Ríos.* Guaranda, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropeciarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Alvarado, C. D. (2018). *Evaluación de incidencia de problemas fitosanitarios en híbrido de café robusta (Coffea canephora Pierre) con cinco densidades de siembra en Cantón Caluma, Provincia Bolívar.* Guaranda, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Anchapaxi, C. R. (2013). *Evaluación agronómica del cultivo de maíz suave en choclo (Zea mays amylacea) variedad Mishca con utilización de tres abonos orgánicos en tres dosis en zona de Pifo, Provincia de Pichincha.* Guaranda, Provincia Bolívar, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Arévalo, T. J. (2012). *Evaluación de sistemas de cultivo, prácticas de labranza y rotación de cultivo de maíz duro (Zea Mays L) en microcuenca del Río Alumbre, Cantón Chillanes, Provincia Bolívar, Ecuador.* Guaranda, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Calero, C. F. (2019). *Evaluación morfológica y agronómica de cacao (Teobroma cacao L) nacional fino de aroma con implementación de diferentes prácticas agrícolas en fertilización y poda.* Guaranda, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Capa, S. H. (2015). *Probabilidades y estadística: Para una gestión científica de la información.* Ladrón de Guevara E11-253, Quito, Ecuador: Unidad de Publicaciones de la Facultad de Ciencias y Vicerrectorado de Docencia de la Escuela Politécnica Nacional.

Carhuaricra, R. J. (2013). *Caracterización morfo agronómica de 10 accesiones de cebada de grano desnudo (Hordeum vulgare L) en Granja Laguacoto II, Cantón Guaranda, Provincia Bolívar.* Guaranda, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Chávez, V. M. (2018). *Validación agronómica de 20 cultivares de fréjol arbustivo (Phaseolus vulgaris L) en Localidad de Monte Carlos, Cantón Urdaneta, Provincia Los Ríos.* Guaranda, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Chicaiza, P. J., & Curichumbi, P. D. (2012). *Estudio de línea base de producción de papa (Solanum tuberosum L) en cinco comunidades de Parroquía La Matriz, Cantón Guamote, Provincia Chimborazo.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Córdova, T. J., & Solís, A. M. (2019). *Evaluación agronómica de respuesta de cuatro líneas promisorias de trigo duro (Triticum durum) a fertilización nitrogenada en dos localidades de Provincia Bolívar.* Guaranda, Provincia de Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Curipallo, S. D. (2013). *Evaluación de cuatro dosis de ácido giberélico en estado fisiológico de floración de tomate de árbol (Cyphornandra betacea) en Patate, Provincia Tungurahua.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Gonzaga, O. D. (2013). *Recuperación de palma aceitera (Elaeis guineensis Jacq) bajo estrés por desbalance catiónico de Ca, Mg y K con uso de diferentes fuentes de Mg y K, La Concordia.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

González, B. G. (1985). *Métodos Estadísticos y Principios de Diseño Experimental.* Quito, Ecuador: Editorial Universitaria. Universidad Central del Ecuador.

González, C. L. (2015). *Identificación de sistemas de producción de maíz suave (Zea mays L) en micro cuenca del Río San Pablo, Cantón San Miguel, Provincia Bolívar.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturalesn y del Ambiente. Universidad Estatal de Bolívar.

Gualotuña, L. E. (2012). *Evaluación agronómica de eficiencia de uso de nitrógeno en cultivo de cebada (Hordeum vulgare L) variedad INIAP-Guaranga 2010 con cinco niveles de fertilización nitrogenada en Granja Laguacoto II, Cantón Guaranda.* Guaranda, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Gujarati, D. N., & Porter, D. C. (2010). *Econometría.* Colonia Desarrollo Santa Fe. Delegación Álvaro Obregón. México, D. F.: McGraw-Hill/Interamericana Editores, S.A. DE C.V.

Juela, M. L., & Mosquera, C. P. (2019). *Evaluación agro morfológica de plántulas de cinco variedades de naranja mediante multiplicación in vitro de embriones inmaduros con tres dosis de ácido giberélico.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Ledesma, J. M. (2017). *Valoración de canales de comercialización del rubro naranja (Citrus sinensis L) en Cantón Caluma, Provincia Bolívar.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturalesn y del Ambiente. Universidad Estatal de Bolívar.

Levine, D. M., Krehbiel, T. C., & Berenson, M. L. (2006). *Estadística para Administración.* Col. Industrial Atoto. Naucalpan de Juárez, estado de México: Pearson Educación de México, S. A. de C. V.

Moya, C. E. (2018). *Valoración de producción y comercialización de maíz duro (Zea Mays L) en Cantón Pueblo Viejo, Provincia Bolívar.* Guarnada, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Ramírez, M. P. (2001). *Principios de Econometría.* Universidad Autónoma Chapingo: Centro de Investigaciones Económicas Sociales y Tecnológicas de la Agroindustria y la Agricultura Mundial (CIESTAAM). Universidad Autónoma Chapingo (UACh).

Remache, P. J. (2012). *Caracterización morfo agronómica de 24 accesiones de trigo duro (Triticum turgidum L, Thell durum) en Localidad Laguacoto II, Cantón Guaranda, Provincia Bolívar.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Rumiguano, Q. M. (2019). *Respuesta agronómica del maíz (Zea mays L) INIAP-111 a fertilización nitrogenada y tres tipos de labranza en Chalongoto, Cantón Guaranda, Provincia Bolívar.* Guarnada, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Suárez, C. L. (2013). *Evaluación de producción de cebada (Hordeum vulgare L) con prácticas agroforestales de conservación del suelo en microcuenca del Río Illangama, Cantón Guaranda, Provincia Bolívar, Ecuador.* Guaranda, Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Ulloa, V. E. (2012). *Caracterización de producción de naranja (Citrus sinensis L) en Parroquia Las Mercedes, Cantón Las Naves, Provincia Bolívar.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Urrutia, Q. S. (2013). *Evaluación del efecto de protectores sobre la producción y calidad del banano en Recinto Pailón, Chacarita, Cantón Ventanas, Provincia Los Ríos.* Guaranda, Ecuador: Escuela de Agronomía. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Váldez, R. J. (2015). *Evaluación morfo agronómica y productiva de ocho variedades de arroz (Oriza sativa L) en Recinto Los Cerritos, Cantón Urdaneta, Provincia Los Ríos.* Guaranda, Ecuador: Escuela de Agronomía. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Vela, V. J. (2013). *Validación de seis formulaciones de fertilizantes químicos en cultivo de papa variedad súper chola (Solanum tuberosum L) en Comunidad de Guitig Alto, Cantón Mejía, Provincia Pichincha.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.

Wackerly, D. D., Mendenhall, W. I., & Scheaffer, R. L. (2010). *Estadística matemática con aplicaciones.* Col. Cruz Manca, Santa Fe. México, D. F.: Cengage Learning Editores, S. A. de C. V., una Compañía de Cengage Learning, Inc.

Yépez, E. A., & Erazo, B. A. (2015). *Estudio de línea base con canales de comercialización del rubro para (Solanum tuberosum L) en dos sectores de Parroquía Guanujo, Cantón Guaranda, Provincia Bolívar.* Guaranda, Provincia Bolívar. Ecuador: Escuela de Ingeniería Agronómica. Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente. Universidad Estatal de Bolívar.